CLASSIFICATION ALGORITHMS WITH DIFFERENTIAL PRIVACY AND FAIRNESS

CLASSIFICATION ALGORITHMS WITH DIFFERENTIAL PRIVACY AND FAIRNESS GUARANTEES

BY

HRAD GHOUKASIAN, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF COMPUTING AND SOFTWARE

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2024) McMaster University

(Computing and Software) Hamilton, Ontario, Canada

TITLE:	Classification Algorithms with Differential Privacy and Fairness Guarantees
AUTHOR:	Hrad Ghoukasian B.Sc. (Electrical Engineering), Sharif University of Technology, Tehran, Iran
SUPERVISOR:	Dr. Shahab Asoodeh
NUMBER OF PAGES:	xiii, 101

Lay Abstract

Fairness and privacy are two key concepts in trustworthy machine learning. In high-stakes scenarios, models must protect individual privacy while avoiding discrimination against demographic subgroups. Differential privacy (DP), the standard notion of privacy in machine learning these days, is divided into two main categories: central DP and local DP (LDP). The first part of this thesis examines the interplay between central DP and fairness in binary classification, presenting an algorithm that guarantees both privacy and fairness while providing theoretical performance guarantees. This algorithm is evaluated on real-world datasets, showing improved fairness without compromising privacy or utility. The second part introduces an optimal data pre-processing method using LDP to minimize unfairness, demonstrating an application of LDP in reducing unfairness in model predictions. Experiments on various datasets show that this optimal pre-processing outperforms existing LDP-based pre-processing fairness intervention methods and state-of-the-art fairness post-processing, achieving better fairness while maintaining comparable utility, even when compared to non-private scenarios.

Abstract

Machine learning algorithms are increasingly used in high-stakes decision-making tasks, highlighting the need to evaluate their trustworthiness, especially regarding privacy and fairness. Models must protect individual privacy and avoid discriminating against demographic subgroups. Differential Privacy (DP) has become the standard for privacy-preserving machine learning. It is generally divided into central DP, which relies on a trusted curator, and local DP (LDP), where no trusted entity is assumed.

The first part of this thesis investigates binary classification under the constraints of both central DP and fairness. We propose an algorithm based on the decoupling technique to learn a classifier that guarantees fairness. This algorithm takes classifiers trained on different demographic groups and produces a single classifier satisfying statistical parity. We then refine this algorithm to incorporate DP. The performance of the resulting algorithm is rigorously analyzed in terms of privacy, fairness, and utility guarantees. Empirical evaluations on the Adult and Credit Card datasets show that our algorithm outperforms state-of-the-art methods in fairness while maintaining the same levels of privacy and utility.

The second part of this thesis addresses the design of an optimal pre-processing

method based on LDP mechanisms to minimize data unfairness and reduce classification unfairness. For binary sensitive attributes, we derive a closed-form expression for the "optimal" mechanism. For non-binary sensitive attributes, we formulate an optimization problem that, when solved algorithmically, yields the optimal mechanism. We theoretically prove that applying these pre-processing mechanisms leads to lower classification unfairness using the notion of discrimination-accuracy optimal classifiers. Empirical evaluations on multiple datasets demonstrate the effectiveness of these mechanisms in reducing classification unfairness, highlighting LDP's potential as a tool for enhancing fairness. This contrasts with central DP, which has been shown to adversely affect fairness.

To my beloved parents, Natalie and Hrier

Acknowledgements

I am deeply grateful to my exceptional supervisor, Dr. Shahab Asoodeh, for his constant support and inspiration. His thoughtful and patient guidance has shaped my research journey, helping me grow both personally and professionally. I have always been inspired by his passion for research and his engaging presentation style, which have deeply influenced my own enthusiasm—I owe Shahab my passion for both research and presentation. Our meetings have always been a highlight of my week, providing motivation and valuable insights that fueled my progress. His encouragement and belief in my potential have been vital to my development, and I am excited for our continued collaboration. Truly, I could not have asked for a more supportive and dedicated advisor.

I am sincerely grateful to Dr. Hassan Ashtiani for his inspiring guidance. His course and our group discussions have been incredibly valuable and insightful.

I wish to express my gratitude to my supervisory committee members, Dr. Hassan Ashtiani and Dr. Sivan Sabato, for reviewing my thesis. Their thoughtful insights and feedback have greatly contributed to my academic growth.

I am thankful to Dr. Xiaoxiao Li and Dr. Sai Praneeth Karimireddy, my mentors and collaborators during my internship. I learned so much from working with them, and I truly appreciate their guidance and support.

I would like to thank my friends Shirak, Mohammad, Alireza, Narges, Behnoosh, Daei, and Hunter for their wonderful friendships and the enriching conversations we have shared. A special thank you to Behnoosh for her continuous guidance throughout my master's.

Finally, I want to express my deepest gratitude to my family, who have been with me from the very beginning. To my parents, Natalie and Hrier, thank you for your endless love, and for providing me with every opportunity to learn and grow—I am profoundly grateful. I also want to thank my grandparents for their constant support and encouragement, especially during my time away from home.

Contents

List of Figures

List of Tables

Author's Declaration

I, Hrad Ghoukasian, declare that this thesis titled "Classification Algorithms with Differential Privacy and Fairness Guarantees" is my own work, developed under the supervision and collaboration of Dr. Shahab Asoodeh. The thesis is composed of two parts: (1) the intersection of fairness with central differential privacy in binary classification, and (2) the intersection of fairness with local differential privacy.

A Part of the work presented in this thesis has been published in our paper (Ghoukasian and Asoodeh, 2024), included in the proceedings of the 2024 IEEE International Symposium on Information Theory. This publication primarily addresses the first part, which explores the intersection of central differential privacy and fairness in classification.

Chapter 1

Introduction

1.1 Motivation

Machine learning algorithms are increasingly deployed in high-stakes decision-making processes, making it crucial to rigorously evaluate their trustworthiness. Two critical aspects of trustworthy machine learning are privacy and fairness. On the one hand, machine learning models must protect the privacy of individuals within training datasets, while on the other hand, they should prevent discrimination against any demographic subgroups.

Differential Privacy (DP) has emerged as the de-facto standard for ensuring privacy in machine learning applications. Informally speaking, a randomized algorithm is considered differentially private if its output distribution does not change significantly when an individual entry in the dataset is altered (i.e., when moving between neighboring datasets) (Dwork *et al.*, 2006; Dwork, 2006). This framework has been widely adopted in practice (e.g., Erlingsson *et al.* (2014); Differential privacy team Apple (2017); Kifer *et al.* (2020); Rogers *et al.* (2020)). DP is generally classified

into two categories: central DP, which assumes the presence of a trusted curator, and local DP (LDP), which operates without such a trusted entity.

The notion of privacy under DP is well-defined. However, unlike privacy, there is no universal definition of fairness. Therefore, how fairness is defined or algorithmically enforced depends on the particular context of the problem. Previous research has shown that central DP can, in some cases, worsen fairness of predictions or even be incompatible with fairness objectives (Bagdasaryan *et al.*, 2019; Ganev *et al.*, 2022; Pujol *et al.*, 2020; Farrand *et al.*, 2020; Agarwal, 2020; Cummings *et al.*, 2019). In contrast, the relationship between LDP and fairness remains underexplored, with no established evidence of incompatibility.

In this thesis, we explore the intersection of DP and fairness, focusing on both central DP and LDP. In this section, we begin by providing precise definitions for both central and local DP and reviewing various fairness metrics commonly used in the literature. We then discuss existing research on fair classification under central DP. Finally, we investigate how LDP can be utilized as a pre-processing mechanism to ensure fairness in classification tasks. These will be the central themes of this thesis.

1.1.1 Differential Privacy

LDP and central DP differ fundamentally in their approaches to privacy preservation. In central DP, the aggregation function, which combines data from the entire dataset to compute a summary statistic (e.g., mean or sum), is executed by a trusted server. The server then applies noise to the output to ensure privacy. This centralized approach assumes the existence of a trusted curator who has access to the unperturbed data. Conversely, LDP does not assume a trusted entity; instead, each data point

is perturbed locally before being sent to the server. This protects each individual's data, even if the server is compromised. LDP ensures privacy by adding noise at the individual data entry level, decentralizing the privacy mechanism. We now formally define these two concepts.

Definition 1. *(Central differential privacy (Dwork et al., 2006; Dwork, 2006)). A randomized mechanism $M : \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ is (ε, δ)-differentially private ((ε, δ)-DP) if for any pair of neighboring datasets D and D' that differ in exactly one record, and for any subsets of outputs $S \subseteq \mathcal{R}$, we have,*

$$\Pr(M(D) \in S) \leq e^{\varepsilon} \Pr(M(D') \in S) + \delta.$$

Definition 2 (Local differential privacy (Warner, 1965; Evfimievski *et al.*, 2003))**.** *A randomized algorithm $M : \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ is said to satisfy ε-local differential privacy (ε-LDP), where $\varepsilon > 0$, if for any pair of input values $x_1, x_2 \in \mathcal{D}$, and for any possible output $y \in \mathcal{R}$, it holds that*

$$\Pr(M(x_1) = y) \leq e^{\varepsilon} \Pr(M(x_2) = y).$$

1.1.2 Fairness

Fairness in machine learning aims to prevent models from discriminating against demographic subgroups, which are typically distinguished by *sensitive* attributes, such as gender or race. However, the definition and enforcement of fairness depend on the specific context of the problem, and no single universal definition exists. Below are the formal definitions of some of the classification group fairness notions used in

this work.

Definition 3 (Equalized opportunity (Hardt *et al.*, 2016)). *Let $X \in \mathcal{X}$, $A \in [k]$, and $Y \in \{0,1\}$ be random variables representing non-sensitive features, sensitive attributes, and labels, respectively, with a joint distribution P_{XAY}. Let $\hat{Y} = \hat{h}(X,A)$ be a binary classifier, where $\hat{h} : \mathcal{X} \times [k] \to \{0,1\}$. The equalized opportunity gap $\Delta_{\mathrm{EO}}(\hat{h})$ associated with the classifier $\hat{h}$ and distribution P_{XAY} is defined as:*

$$\Delta_{\mathrm{EO}}(\hat{h}) = \max_{a,a' \in [k]} \left| \Pr(\hat{Y} = 1 \mid A = a, Y = 1) - \Pr(\hat{Y} = 1 \mid A = a', Y = 1) \right|.$$

Definition 4 (Statistical parity (Feldman *et al.*, 2015)). *Let $X \in \mathcal{X}$, $A \in [k]$, and $Y \in \{0,1\}$ be random variables representing non-sensitive features, sensitive attributes, and labels, respectively, with a joint distribution P_{XAY}. Let $\hat{Y} = \hat{h}(X,A)$ be a binary classifier, where $\hat{h} : \mathcal{X} \times [k] \to \{0,1\}$. The statistical parity gap $\Delta_{\mathrm{SP}}(\hat{h})$ associated with the classifier $\hat{h}$ and distribution P_{XAY} is defined as:*

$$\Delta_{\mathrm{SP}}(\hat{h}) = \max_{a,a' \in [k]} \left| \Pr(\hat{Y} = 1 \mid A = a) - \Pr(\hat{Y} = 1 \mid A = a') \right|.$$

We say that $\hat{Y} = \hat{h}(X,A)$ satisfies γ-statistical parity if

$$\Delta_{\mathrm{SP}}(\hat{h}) \leq \gamma.$$

Definition 5 (Mean Equalized Odds (Hardt *et al.*, 2016; Alghamdi *et al.*, 2022)). *Let $X \in \mathcal{X}$, $A \in [k]$, and $Y \in \{0,1\}$ be random variables representing non-sensitive features, sensitive attributes, and labels, respectively, with a joint distribution P_{XAY}. Let $\hat{Y} = \hat{h}(X,A)$ be a binary classifier, where $\hat{h} : \mathcal{X} \times [k] \to \{0,1\}$. The mean*

equalized odds gap $\Delta_{\text{MEO}}(\hat{h})$ *associated with the classifier* $\hat{h}$ *and distribution* P_{XAY} *is defined as:*

$$\Delta_{\text{MEO}}(\hat{h}) = \max_{a,a' \in [k]} \frac{1}{2}\Big(\big|\text{TPR}_{A=a} - \text{TPR}_{A=a'}\big| + \big|\text{FPR}_{A=a} - \text{FPR}_{A=a'}\big|\Big),$$

where $\text{TPR}_{A=a} = \Pr(\hat{Y} = 1 \mid Y = 1, A = a)$ *and* $\text{FPR}_{A=a} = \Pr(\hat{Y} = 1 \mid \ Y = 0, \ A = a)$.

It is worth noting that definitions of classification unfairness extend beyond the three mentioned here. Numerous definitions exist (Chouldechova, 2017; Dwork *et al.*, 2012; Berk *et al.*, 2021; Corbett-Davies *et al.*, 2017; Kilbertus *et al.*, 2017; Kleinberg *et al.*, 2016; Kusner *et al.*, 2017; Sabato *et al.*, 2024) that may be useful to consider depending on the specific context of the problem.

1.1.3 Central DP and Fairness in Classification

In the intersection of fairness and privacy in classification, one line of work explores the relationship between DP and different notions of fairness, examining their compatibility and determining how fairness affects DP guarantees and vice versa (Bagdasaryan *et al.*, 2019; Cummings *et al.*, 2019; Mangold *et al.*, 2023; Chang and Shokri, 2021a; Pujol *et al.*, 2020; Farrand *et al.*, 2020; Agarwal, 2020). Another research direction focuses on developing classification algorithms that simultaneously guarantee DP and a specific notion of fairness. Despite these significant advances, most of these approaches have inherent limitations. For instance, some methods are designed specifically for certain types of classification models, such as logistic regression, and are not applicable to other models (Xu *et al.*, 2019; Jagielski *et al.*, 2019; Ding *et al.*,

2020). Other studies focus solely on protecting the privacy of sensitive attributes, neglecting the privacy of other features. Specifically, while some methods ensure that sensitive attributes remain confidential, they do not provide the same privacy guarantees for non-sensitive attributes, leaving them susceptible to leakage (Jagielski *et al.*, 2019; Mozannar *et al.*, 2020; Tran *et al.*, 2021a, 2022). Finally, although many of these methods have demonstrated practical effectiveness, they often lack theoretical guarantees regarding utility and the extent of fairness violations (Xu *et al.*, 2019; Ding *et al.*, 2020; Xu *et al.*, 2021; Tran *et al.*, 2021b,a, 2022; Esipova *et al.*, 2023; Lowy *et al.*, 2023; Yaghini *et al.*, 2023). Our goal is to develop a binary classification algorithm with provable DP and fairness guarantees that addresses these limitations in DP-Fair classification algorithms.

1.1.4 LDP as a Pre-Processing Tool for Fairness

Although some progress has been made at the intersection of LDP and fairness, several challenges persist. For instance, Mozannar *et al.* (2020) presents a learning scheme for training non-discriminatory classifiers when only a privatized version of the sensitive attribute is available. Specifically, the sensitive attribute is privatized using an ε-LDP mechanism, generalized randomized response (GRR) (Kairouz *et al.*, 2014), and the approach provides theoretical performance guarantees. Similarly, Chen *et al.* (2022) explores a "semi-private" setting where a small subset of users share their sensitive attributes without modification, while the remaining users employ an ε-LDP protocol. While these frameworks provide approaches for fair prediction with LDP-privatized data, they either focus exclusively on binary sensitive attributes or limit themselves to using GRR as the LDP mechanism, highlighting the need for further exploration.

Recently, some works have specifically addressed the impact of LDP on fairness in classification (Arcolezi *et al.*, 2023; Makhlouf *et al.*, 2024b,a). For instance, Arcolezi *et al.* (2023) empirically examined how applying different LDP mechanisms influences fairness and utility in binary classification. Their results showed that applying LDP to sensitive attributes improved classification fairness with minimal impact on utility compared to training on non-private data. Building on this, Makhlouf *et al.* (2024a) demonstrated that increasing privacy (i.e., lowering ε) further enhances fairness, and applying LDP to multiple sensitive attributes more effectively reduces disparity than focusing on a single attribute. Additionally, Makhlouf *et al.* (2024b) provided a theoretical analysis of how randomized response (RR) (Warner, 1965), a widely used LDP mechanism, affects fairness by considering privacy levels and data distribution. This work, based on assumptions about the unfairness metric and the learning algorithm used, explored conditions under which privacy reduces or increases unfairness. These findings challenge the common belief in the central DP context, where increased privacy is often thought to worsen fairness. The results for LDP, however, point toward a promising direction: identifying optimal ways to perturb data using LDP mechanisms for both binary and non-binary sensitive attributes. In this work, we explore LDP as a pre-processing method aimed at enhancing fairness in classification tasks.

1.2 Summary of Contributions

The contributions of this thesis are twofold, focusing on the intersections of central DP and fairness in the first part, and LDP and fairness in the second part. More specifically, the contributions are as follows:

1.2.1 Contributions: Central DP and Fairness

- We establish a lower bound for the sum of prediction changes across a pair of subgroups under statistical parity (without privacy) and propose an algorithm (based on (Zhao and Gordon, 2022, Algorithm 1)) that attains this bound (Algorithm 1). The theoretical guarantee of this algorithm is provided in Theorem 7.

- We then propose a differentially private version (Algorithm 2) of Algorithm 1 and derive its theoretical guarantees for fairness and utility — both with high probability and in expectation — in Theorem 8 and Proposition 9.

- Through several experiments on two well-known datasets (Adult and Credit Card), we empirically demonstrate that Algorithm 2 achieves competitive accuracy when compared to the state-of-the-art DP-Fair classification method, DP-FERMI (Lowy *et al.*, 2023). In particular, we show that for a given level of accuracy and privacy, our algorithm provides a significantly better fairness guarantee across both datasets.[1]

1.2.2 Contributions: Local DP and Fairness

- We introduce the problem of designing LDP mechanisms that minimize data unfairness for both binary and non-binary sensitive attribute cases. For the binary case, we provide a closed-form expression of the mechanism that minimizes unfairness while maintaining non-trivial utility. For the non-binary case, we reformulate the optimization problem as a min-max linear fractional program, solvable numerically using the branch-and-bound technique (Jiao and Li,

[1]The experimental code for these experiments can be accessed at https://github.com/hradghoukasian/dp_fair_binary.

2022).

- After identifying the optimal LDP mechanism to reduce data unfairness, we theoretically prove that if the classifier belongs to a certain set of discrimination-accuracy optimal classifiers, training on less discriminatory data will lead to reduced post-classification unfairness. This result justifies the objective function we use to minimize data unfairness.

- Through experiments on various datasets and fairness metrics, we demonstrate that our approach effectively reduces unfairness. We show that it achieves utility metrics comparable to well-known LDP mechanisms while ensuring lower unfairness after classification. Moreover, we compare our optimal mechanism with Fair Projection (Alghamdi *et al.*, 2022), a state-of-the-art post-processing fairness intervention framework (Wang *et al.*, 2024; Denis *et al.*, 2024). Our results demonstrate that, while Fair Projection provides a wide range of accuracy-fairness trade-offs, our optimal mechanism achieves a solution that, for fixed fairness metrics, attains slightly better accuracy or, in the worst case, matches a point on Fair Projection's trade-off curve with significantly less runtime compared to Fair Projection.

1.3 Thesis Organization

The first part of this thesis addresses the intersection of central DP and fairness in classification. Chapter 2 provides background knowledge and a literature review on the intersection of DP and fairness in classification. Section 2.1 discusses decoupled classifiers and stratification, two key concepts in this work. Sections 2.2 and 2.3

give an overview of related work on central DP and fairness in classification, highlighting gaps in the literature. Section 2.4 provides a brief overview of our approach to developing a DP post-processing algorithm that ensures both fairness and utility guarantees, which is thoroughly explored in Chapter 3.

Chapter 3 presents the main results of the first part of this thesis. After defining the notation and the main objectives in Sections 3.1 and 3.2, the main theoretical results are provided in Sections 3.3 and 3.4. Section 3.3 discusses the optimal post-processing algorithm without considering privacy, including its theoretical guarantees. Following this, Section 3.4 introduces the DP version of this algorithm (Algorithm 2), along with the theoretical guarantees for this classifier. Finally, Section 3.5 provides empirical results for Algorithm 2, comparing it with state-of-the-art DP-Fair classifiers.

The second part of this thesis focuses on the intersection of LDP and fairness. Chapter 4 provides a literature review on the intersection of LDP and fairness, fairness intervention methods, and pre-processing techniques for reducing unfairness in classification. These are detailed in Sections 4.1, 4.2, and 4.3. Section 4.4 offers a brief overview of our approach, which leverages LDP as a pre-processing method for achieving fairness in classification. The main results of the second part of the thesis are presented in Chapter 5.

After defining the notation and the preliminary knowledge required for the main results in Sections 5.1 and 5.2, we examine the impact of a specific LDP mechanism, GRR, on data fairness in Section 5.3. Then, in Sections 5.4 and 5.5, we formulate the problem of an optimal LDP-based pre-processing technique for data fairness, considering both binary and non-binary sensitive attributes. For the binary case, the

optimal closed-form mechanism is derived in Section 5.4, while for the non-binary case, the optimization problem is reformulated for numerical solution, as explained in Section 5.5. Finally, the theoretical results are concluded in Section 5.6, where we demonstrate that data fairness leads to classification fairness when the classifier belongs to a specific class. Section 5.7 provides experimental results for the optimal mechanisms presented in Sections 5.4 and 5.5.

Lastly, Chapter 6 summarizes the work in this thesis, discussing limitations and directions for future research. Proofs for the first part of the thesis, covering central DP and fairness, are provided in Appendix A, while proofs for the second part, addressing LDP and fairness, can be found in Appendix B.

Chapter 2

Central DP and Fairness in Classification: Key Concepts and Related Work

A widely recognized concept of fairness is statistical parity (Feldman *et al.*, 2015), also known as demographic parity. This definition implies that the predictions of a classification model should be independent of the *sensitive* attributes, such as gender or race. While privacy and fairness have been extensively studied separately in the literature, their intersection has recently gained attention.

2.1 Decoupled Classifiers and Stratification

In the literature on fair classification algorithms, incorporating sensitive attributes is known as "fairness through awareness" (Dwork *et al.*, 2012), while deliberately omitting sensitive attributes is referred to as "fairness through blindness" or "fairness

through unawareness." It is well-known that simply excluding sensitive attributes does not guarantee fairness, mainly because these attributes can be closely correlated with other data features and provide important context for interpreting the rest of the data (Dwork *et al.*, 2018).

While the decision to use or omit sensitive attributes is problem-specific, Dwork *et al.* (2018) demonstrated that in scenarios where it is legally and ethically permissible to use these attributes, training separate classifiers (decoupled classifiers) for each group can outperform a single optimal classifier in terms of both accuracy and fairness (without privacy considerations). Furthermore, Wang *et al.* (2021) showed that employing decoupled classifiers does not negatively impact any group in terms of average performance metrics within the information-theoretic regime where the underlying data distribution is known.

Building on the concept of decoupled classification combined with differential privacy, Rosenblatt *et al.* (2023) proposed a "stratification" technique to reduce the disparity in differentially private mechanisms. This technique involves applying a DP mechanism separately to different subgroups and then recombining the respective results to derive overall statistics for the entire dataset.

It was demonstrated that a naive stratification approach can produce highly accurate estimates for population-level statistics without requiring an additional privacy budget. Building on this foundation, our pipeline first applies DP and subsequently addresses fairness through a post-processing step. This order of application is crucial because DP methods introduce random noise, which can alter model predictions. If the model met fairness standards before applying DP, the introduction of noise could cause these standards to no longer be met.

2.2 Central DP Negatively Impacts Fairness

A wide range of fairness intervention techniques exist for classification (see Section 4.2). However, a significant challenge arises when privacy and fairness are considered together in classification settings. It has been empirically demonstrated that differentially private mechanisms can exacerbate unfairness (Bagdasaryan *et al.*, 2019; Pujol *et al.*, 2020). This necessitates the implementation of specific strategies to mitigate the negative effects of DP on fairness guarantees.

Bagdasaryan *et al.* (2019) empirically shows that central DP, specifically differentially private stochastic gradient descent (DP-SGD) (Abadi *et al.*, 2016), has a disparate impact on the accuracy of different subgroups. Similarly, Ganev *et al.* (2022) demonstrates that applying DP to synthetic data generation disproportionately affects minority sub-populations. Pujol *et al.* (2020) also finds that applying DP can exacerbate unfairness in decision-making tasks. Farrand *et al.* (2020) further shows that even with loose privacy guarantees and minimal data imbalance, DP-SGD can lead to disparate impact. Agarwal (2020); Cummings *et al.* (2019) introduce an incompatibility result, highlighting that $(\varepsilon, 0)$-DP and fairness are at odds when aiming for non-trivial accuracy in learning algorithms.

Additionally, some works explore the intersection of privacy and fairness beyond the scope of central DP or group fairness. For instance, Chang and Shokri (2021b) examines the privacy risks of achieving group fairness, showing that fairness may compromise privacy, particularly through membership inference attacks, even when DP is not explicitly applied. Dwork *et al.* (2012) provides a theoretical link between individual fairness and DP, arguing that individual fairness can be seen as a generalization of DP and outlining conditions under which a DP mechanism ensures

individual fairness. A recent study (Ko *et al.*, 2024) examined the impact of DP on fairness beyond classification tasks. They demonstrated that using DP to set sampling rates for collecting socio-demographic data reduces unfairness across different population segments in resource allocation.

2.3 DP-Fair Classification Algorithms

Another research direction at the intersection of privacy and fairness focuses on developing classification algorithms that simultaneously ensure DP and a specific notion of fairness.

Some of these algorithms aim to make existing private algorithms fair. For example, Esipova *et al.* (2023) investigates the causes of unfairness in DP-SGD, identifying gradient misalignment as the primary source of disparate impact. They show that the discrepancy between the direction of unclipped and clipped gradients can create imbalances in gradient norms across groups. Their work proposes a modification to DP-SGD that mitigates this gradient misalignment. Similarly, Yaghini *et al.* (2023) present adjustments to DP-SGD and PATE (Papernot *et al.*, 2018), two privacy-preserving learning approaches, to enhance fairness. Tran *et al.* (2021b) addresses disparate impact in DP-SGD by adding fairness constraints to its objective function. Additionally, Xu *et al.* (2021) proposes a modified DP-SGD algorithm, DP-SGD-F, which adaptively adjusts sample contributions based on group-specific clipping biases to minimize disparate impact on group accuracy.

Other research focuses on making existing fair algorithms private. For instance, Tran *et al.* (2021a) makes the Fair-Lagrangian Dual algorithm private by adding noise to both primal and dual updates. Jagielski *et al.* (2019) introduces a private

implementation of the equalized odds post-processing approach by Hardt *et al.* (2016) and the in-processing reduction approach by Agarwal *et al.* (2018).

Some studies target DP-Fair classifiers for specific models, such as logistic regression. Xu *et al.* (2019) and Ding *et al.* (2020) use the functional mechanism (Zhang *et al.*, 2012), a DP-based method for optimization-based models that perturbs the objective function to maintain privacy.

Lastly, Mozannar *et al.* (2020) introduces two differentially private fair classification algorithms that provide privacy guarantees only for sensitive attributes, using randomized response mechanism that implements the local version of DP. However, their approach does not naturally extend to providing privacy guarantees for all features, especially when dealing with continuous, high-dimensional non-sensitive attributes.

As discussed in Section 1.1.3, despite the comprehensive range of existing methods in DP-Fair classification, most of these approaches still face inherent limitations. They are either restricted to specific classifier structures, fail to provide privacy guarantees for all data features, or lack theoretical guarantees for both fairness constraints and classifier utility. A summary of the characteristics of these methods can be found in Table 2.1.[1]

2.4 Overview of Our Approach

Our objective is to develop a binary classification algorithm with provable DP and fairness guarantees that addresses the limitations highlighted in Table 2.1. Inspired

[1]In this work, we allow access to the sensitive attributes at test time, similar to the approach in Jagielski *et al.* (2019), Mozannar *et al.* (2020), and Esipova *et al.* (2023).

Table 2.1: Benchmark methods in DP-Fair classification

Reference	Applicable to any model	Theoretical guarantee	Privacy w.r.t. all features
Xu *et al.* (2019)	✗	✗	✓
Jagielski *et al.* (2019) (post-proc.)	✓	✓	✗
Jagielski *et al.* (2019) (in-proc.)	✗	✓	✓
Ding *et al.* (2020)	✗	✗	✓
Mozannar *et al.* (2020)	✓	✓	✗
Xu *et al.* (2021)	✓	✗	✓
Tran *et al.* (2021a)	✓	✗	✗
Tran *et al.* (2021b)	✓	✗	✓
Tran *et al.* (2022)	✓	✗	✗
Esipova *et al.* (2023)	✓	✗	✓
Lowy *et al.* (2023)	✓	✗	✓
Yaghini *et al.* (2023) (FairDPSGD)	✓	✗	✓
Yaghini *et al.* (2023) (FairPATE)	✓	✗	✓
This Work	✓	✓	✓

by the success of decoupled classifiers and the effectiveness of stratification in reducing disparate impact in differentially private mechanisms (as noted in Wang *et al.* (2021); Dwork *et al.* (2018); Rosenblatt *et al.* (2023)), our approach begins with separate classifiers for each sub-population. These classifiers then go into a post-processing step to generate a single classifier. Following the method used in Jiang *et al.* (2020), our goal is to apply a post-processing technique that achieves statistical parity, ensuring that only minimal changes are made to the predictions of the original classifiers. We begin by slightly modifying the approach described in Zhao and Gordon (2022), initially

developed for non-private settings (Algorithm 1). Then, we introduce Algorithm 2, a new method for binary classification that is both differentially private and fair. This new algorithm comes with theoretical guarantees for utility, fairness, and DP.

Chapter 3

Central DP and Fairness in Classification: Main Results

3.1 Notation

In this section, we consider a binary classification setting where there is a joint distribution μ over the triplet $T = (X, A, Y)$, where $X \in \mathcal{X} \subset \mathbb{R}^d$ is the feature vector of non-sensitive attributes, $A \in \{0, 1\}$ is the sensitive attribute, and $Y \in \{0, 1\}$ is the target output. We use $\mu(Y)$ to denote the marginal distribution of Y from a joint distribution μ over Y and some other random variables. Denote the marginal distribution of input X by μ^X. For $a \in \{0, 1\}$, we use μ_a to denote the conditional distribution of (X, Y) conditioned on $A = a$, and μ_a^X to mean the marginal distribution of input X given $A = a$. For any group-aware classifier $h : \mathcal{X} \times \{0, 1\} \to \{0, 1\}$ and $a \in \{0, 1\}$, we also use $h_a(\cdot)$ to denote the restriction of h on $A = a$, respectively, i.e., $h_a(\cdot) := h(\cdot, a)$. Finally, the probabilistic inequality $X \leq_\eta Y$ for a pair of random variables (X, Y) denotes the mathematical statement that $\mathbb{P}(X > Y) \leq \eta$.

3.2 Main Objectives

In this chapter, we develop a framework for designing a binary classification algorithm with provable DP and fairness guarantees. Following Zhao and Gordon (2022); Xu *et al.* (2019); Yaghini *et al.* (2023); Ding *et al.* (2020), we adopt statistical parity as a metric for fairness. We begin in Section 3.3 by detailing our approach to achieve fairness and the specified utility target in a non-private setting. Here, as outlined in Section 2.4, our framework involves partitioning the dataset according to sensitive attributes. We first develop a separate classifier for each subgroup. We then implement a post-processing technique that carefully combines those decoupled classifiers in a way to achieve statistical parity, with the objective of minimally perturbing the original classifiers' predictions. Then in Section 3.4, we expand this framework to include DP.

We consider two classifiers, $h_0^* : \mathcal{X} \to \{0,1\}$ and $h_1^* : \mathcal{X} \to \{0,1\}$, each trained on subgroups specified by the sensitive attribute A with values 0 and 1, respectively. These classifiers are designed to maximize accuracy without initially considering fairness constraints. In a non-private setting, h_0^* and h_1^* are standard classifiers, whereas in private settings, they are considered classifiers learned by DP guarantees. Our post-processing method aims to derive a fair classifier $\hat{h} : \mathcal{X} \times \{0,1\} \to \{0,1\}$ from h_0^* and h_1^*. Following the methodology of Jiang *et al.* (2020), our objective is to achieve this goal by minimally perturbing the predictions of the original classifiers. We thus seek fair $\hat{h}$ that optimizes the utility measured in terms of the sum $\mathbb{P}_{\mu_0^X}(\hat{h}_0(X) \neq h_0^*(X)) + \mathbb{P}_{\mu_1^X}(\hat{h}_1(X) \neq h_1^*(X))$. The reason for adopting the prediction changes over μ_0^X and μ_1^X (as opposed to the combined distribution μ^X) is as follows: when the demographic subgroups are imbalanced in the overall population, relying

only on prediction changes across the combined distribution μ^X can be misleading. This approach may hide significant prediction shifts of the less-represented group, which would be more apparent if we examined the sum of the prediction changes within each subgroup's distribution (μ_0^X and μ_1^X). In other words, it might be possible to have small overall prediction changes across the combined distribution μ^X while minority groups are experiencing substantial changes in their predictions (Zhao and Gordon, 2022).

In Section 3.3, we explore the non-private scenario, discussing Algorithm 1 that ensures statistical parity and optimal utility. In Section 3.4, we extend the algorithm to take privacy into consideration. In particular, we propose Algorithm 2, which outputs a classifier that guarantees DP and achieves statistical parity while maintaining minimal changes in the predictions of the original classifiers, both in expectation and with high probability. To discuss our utility metric under the constraint of statistical parity, we rely on the following proposition.

Proposition 6. *Let $h_0^* : \mathcal{X} \to \{0,1\}$ and $h_1^* : \mathcal{X} \to \{0,1\}$, be arbitrary classifiers trained on subgroups specified by the sensitive attribute $A = 0$ and $A = 1$, respectively. If a predictor $\hat{Y} = \hat{h}(X, A)$ satisfies γ-statistical parity, then*

$$\mathbb{P}_{\mu_0^X}(\hat{h}_0(X) \neq h_0^*(X)) + \mathbb{P}_{\mu_1^X}(\hat{h}_1(X) \neq h_1^*(X)) \geq \left|\mathbb{P}_{\mu_0^X}(h_0^*(X) = 1) - \mathbb{P}_{\mu_1^X}(h_1^*(X) = 1)\right| - \gamma.$$

3.3 Fair Post-Processing without Privacy

Let $h_0^* : \mathcal{X} \to \{0,1\}$ and $h_1^* : \mathcal{X} \to \{0,1\}$ be decoupled classifiers, that is they are trained on subgroups associated with the sensitive attribute $A = 0$ and $A = 1$, respectively. h_0^* and h_1^* can be any arbitrary classifiers. In the context of our analysis, they can be considered as the best available classifiers trained on μ_0 and μ_1. Our goal is to develop an optimal fair classifier $\hat{h} : \mathcal{X} \times \{0,1\} \to \{0,1\}$ using h_0^* and h_1^*. More precisely, we seek $\hat{h}$ that satisfies $\Delta_{SP}(\hat{h}) = 0$ while attaining the lower bound in Proposition 6 for $\gamma = 0$:

$$\mathbb{P}_{\mu_0^X}(\hat{h}_0(X) \neq h_0^*(X)) + \mathbb{P}_{\mu_1^X}(\hat{h}_1(X) \neq h_1^*(X)) = \left|\mathbb{P}_{\mu_0^X}(h_0^*(X) = 1) - \mathbb{P}_{\mu_1^X}(h_1^*(X) = 1)\right|.$$

We present Algorithm 1 to address this objective. This algorithm, which is a slightly modified version of the algorithm in Zhao and Gordon (2022), constructs the fair classifier h^*_{Fair}. The original algorithm in Zhao and Gordon (2022) builds a fair optimal classifier assuming oracle access to the Bayes optimal classifiers h_0^* and h_1^*. Our ultimate goal, to be discussed in the next section with Algorithm 2, is to identify a classifier that is both private and fair. However, the assumption of having access to Bayes optimal classifiers is not practical in DP settings, primarily due to the necessity of introducing noise. Consequently, in Algorithm 1, h_0^* and h_1^* are considered to be any arbitrary classifiers trained on subgroups specified by the sensitive attribute $A = 0$ and $A = 1$ (not necessarily the Bayes optimal classifiers).

Theorem 7. *The classifier h^*_{Fair} constructed by Algorithm 1 satisfies perfect statistical parity ($\Delta_{SP}(h^*_{Fair}) = 0$) and is optimal in terms of the sum of prediction changes*

Algorithm 1 Optimal Fair Binary Classifier

Input: Classifiers h_0^* and h_1^*

Output: A randomized classifier $h_{\text{Fair}}^* : \mathcal{X} \times \{0,1\} \to \{0,1\}$

1: Compute $\alpha = \mathbb{P}_{\mu_0^X}(h_0^*(X) = 1)$ and $\beta = \mathbb{P}_{\mu_1^X}(h_1^*(X) = 1)$. W.L.O.G. assume $\alpha \geq \beta$

2: For (x, a), randomly sample s from the uniform distribution $U(0,1)$

3: Construct h_{Fair}^* as follows:

$$h_{\text{Fair}}^*(x,a) := \begin{cases} a = 0: & \begin{cases} 0 & \text{if } h_0^*(x) = 0 \text{ or } h_0^*(x) = 1 \text{ and } s > \frac{\alpha+\beta}{2\alpha} \\ 1 & \text{if } h_0^*(x) = 1 \text{ and } s \leq \frac{\alpha+\beta}{2\alpha} \end{cases} \\ a = 1: & \begin{cases} 0 & \text{if } h_1^*(x) = 0 \text{ and } s > \frac{\alpha-\beta}{2(1-\beta)} \\ 1 & \text{if } h_1^*(x) = 1 \text{ or } h_1^*(x) = 0 \text{ and } s \leq \frac{\alpha-\beta}{2(1-\beta)} \end{cases} \end{cases}$$

return h_{Fair}^*

compared to classifiers h_0^* and h_1^*, that is

$$\mathbb{P}_{\mu_0^X}(h_{Fair0}^*(X) \neq h_0^*(X)) + \mathbb{P}_{\mu_1^X}(h_{Fair1}^*(X) \neq h_1^*(X)) = \left| \mathbb{P}_{\mu_0^X}(h_0^*(X) = 1) - \mathbb{P}_{\mu_1^X}(h_1^*(X) = 1) \right|.$$

3.4 Fair Post-Processing with Privacy

Next, we delve into a private version of Algorithm 1. It is worth noting that achieving $(\varepsilon, 0)$-DP is incompatible with non-trivial guarantees of fairness and utility (see, e.g., Cummings *et al.* (2019); Agarwal (2020) for more details). As a result, our objective is to design an (ε, δ)-DP version of Algorithm 1 with comparable fairness and utility guarantees, provided that $\delta > 0$.

First, notice that this goal is not feasible just by replacing the input classifiers h_0^* and h_1^* with some differentially private decoupled classifiers that are learned by applying an existing differentially private learning method —most notably, DP-SGD— to each demographic subgroup. This approach implicitly assumes privacy guarantees

only for non-sensitive features, whereas we aim to guarantee privacy for both sensitive and non-sensitive features. Additionally, note that Algorithm 1 involves some computations over the dataset (e.g., in computing α and β in line 1). This therefore necessitates some changes in Algorithm 1 for constructing its differentially private version. Let $h^*_{\varepsilon,\delta} : \mathcal{X} \times \{0,1\} \to \{0,1\}$ be a classifier guaranteeing (ε, δ)-DP with respect to all features. Then, $h^*_{\varepsilon,\delta,0} : \mathcal{X} \to \{0,1\}$ and $h^*_{\varepsilon,\delta,1} : \mathcal{X} \to \{0,1\}$ represent the restrictions of $h^*_{\varepsilon,\delta}$ to $A = 0$ and $A = 1$, respectively. Thus, our goal can be formulated as follows: Given classifiers $h^*_{\varepsilon,\delta,0}$ and $h^*_{\varepsilon,\delta,1}$ that are (ε, δ)-DP, we wish to generate a fair classifier $h^*_{\varepsilon',\delta',\text{Fair}}$ with the following properties:

1. $h^*_{\varepsilon',\delta',\text{Fair}}$ is (ε', δ')-DP with some ε' and δ' (depending on ε and δ),

2. $h^*_{\varepsilon',\delta',\text{Fair}}$ satisfies γ-statistical parity with γ being a positive value close to zero with high probability and in expectation,

3. Among all classifiers satisfying the same level of statistical parity gap as $h^*_{\varepsilon',\delta',\text{Fair}}$, the classifier $h^*_{\varepsilon',\delta',\text{Fair}}$ performs comparably with the optimal classifier in terms of utility, both with high probability and in expectation. Optimality here is defined based on minimizing the total number of prediction changes across the distributions μ_0^X and μ_1^X, relative to the predictions made by $h^*_{\varepsilon,\delta,0}$ and $h^*_{\varepsilon,\delta,1}$.

To this goal, we present Algorithm 2 for learning such $h^*_{\varepsilon',\delta',\text{Fair}}$, assuming that the number of data points belonging to each subgroup in a dataset is publicly known. For any dataset D, θ denotes the proportion of data points with $A = 0$, while $\bar{\theta} = 1 - \theta$ represents the proportion of data points with $A = 1$. Recall that Algorithm 1 requires estimates of the proportions of data points in each subgroup classified as label one (denoted by α and β). Algorithm 2 is designed to privately estimate these quantities

Algorithm 2 Private and Fair Binary Classifier with Utility Gap Guarantee

Input: Classifiers $h^*_{\varepsilon,\delta,0}$ and $h^*_{\varepsilon,\delta,1}$, dataset $D = (X_i, A_i, Y_i)_{i=1}^n$ with θn individuals where $A_i = 0$ and $\bar{\theta} n$ individuals where $A_i = 1$, privacy parameters ε_0 and ε_1

Output: Classifier $h^*_{\varepsilon',\delta',\text{Fair}} : \mathcal{X} \times \{0,1\} \to \{0,1\}$ with $\varepsilon' = \varepsilon + \varepsilon_0 + \varepsilon_1$ and $\delta' = \delta$

1: Find $\bar{\alpha} = \frac{1}{n\theta} \sum_{\substack{i=1 \\ A_i=0}}^{n} h^*_{\varepsilon,\delta,0}(X_i)$ and $\bar{\beta} = \frac{1}{n\theta} \sum_{\substack{i=1 \\ A_i=1}}^{n} h^*_{\varepsilon,\delta,1}(X_i)$

2: Add noise to $\bar{\alpha}$ and $\bar{\beta}$:
Sample l_0 from $\text{Lap}\big(\frac{1}{n\theta\varepsilon_0}\big)$ and l_1 from $\text{Lap}\big(\frac{1}{n\bar{\theta}\varepsilon_1}\big)$. Define $\tilde{\alpha} = [\bar{\alpha} + l_0]_0^1$ and $\tilde{\beta} = \big[\bar{\beta} + l_1\big]_0^1$, where $[\cdot]_0^1$ denotes the projection onto $[0,1]$.

3: For (x, a), randomly sample s from the uniform distribution $U(0,1)$

4: Construct $h^*_{\varepsilon',\delta',\text{Fair}}$ as follows:
if $\tilde{\alpha} \geq \tilde{\beta}$, then

$$h^*_{\varepsilon',\delta',\text{Fair}}(x,a) := \begin{cases} \mathbb{1}[h^*_{\varepsilon,\delta,0}(x) = 1]\mathbb{1}[s \leq \frac{\tilde{\alpha}+\tilde{\beta}}{2\tilde{\alpha}}] & \text{if } a = 0 \\ \mathbb{1}[h^*_{\varepsilon,\delta,1}(x) = 0]\mathbb{1}[s \leq \frac{\tilde{\alpha}-\tilde{\beta}}{2(1-\tilde{\beta})}] + \mathbb{1}[h^*_{\varepsilon,\delta,1}(x) = 1] & \text{if } a = 1 \end{cases}$$

if $\tilde{\alpha} < \tilde{\beta}$, then

$$h^*_{\varepsilon',\delta',\text{Fair}}(x,a) := \begin{cases} \mathbb{1}[h^*_{\varepsilon,\delta,1}(x) = 1]\mathbb{1}[s \leq \frac{\tilde{\alpha}+\tilde{\beta}}{2\tilde{\beta}}] & \text{if } a = 1 \\ \mathbb{1}[h^*_{\varepsilon,\delta,0}(x) = 0]\mathbb{1}[s \leq \frac{\tilde{\beta}-\tilde{\alpha}}{2(1-\tilde{\alpha})}] + \mathbb{1}[h^*_{\varepsilon,\delta,0}(x) = 1] & \text{if } a = 0 \end{cases}$$

return $h^*_{\varepsilon',\delta',\text{Fair}}$

by employing the Laplace mechanism, whose privacy guarantee is well-understood. The following theorem delineates the performance of Algorithm 2 in terms of its achievable privacy and fairness guarantees as well as bounds on its utility.

Theorem 8. *The classifier $h^*_{\varepsilon',\delta',Fair} : \mathcal{X} \times \{0,1\} \to \{0,1\}$ constructed by Algorithm 2 satisfies the following three properties:*

- *(Privacy guarantee) $h^*_{\varepsilon',\delta',Fair}$ satisfies (ε', δ')-DP with $\varepsilon' = \varepsilon + \varepsilon_0 + \varepsilon_1$ and $\delta' = \delta$,*

- *(Fairness guarantee) We have:*

$$\Delta_{SP}(h^*_{\varepsilon',\delta',Fair}) \leq_\eta \frac{\log(4/\eta)}{n\theta\varepsilon_0} + \frac{\log(4/\eta)}{n\bar{\theta}\varepsilon_1} + \sqrt{\log(\frac{8}{\eta})\frac{1}{2n\theta}} + \sqrt{\log(\frac{8}{\eta})\frac{1}{2n\bar{\theta}}},$$

- *(Utility guarantee) Let* $err^*(h^*_{\varepsilon,\delta,0}, h^*_{\varepsilon,\delta,1})$ *be defined as:*

$$\min_{\substack{\hat{h}:\mathcal{X}\times\{0,1\}\to\{0,1\} \\ \Delta_{SP}(\hat{h})\leq\Delta_{SP}(h^*_{\varepsilon',\delta',Fair})}} \left[\mathbb{P}_{\mu_0^X}(\hat{h}_0(X) \neq h^*_{\varepsilon,\delta,0}(X)) + \mathbb{P}_{\mu_1^X}(\hat{h}_1(X) \neq h^*_{\varepsilon,\delta,1}(X))\right]. \tag{3.4.1}$$

Then, we have:

$$\begin{aligned}
&\mathbb{P}_{\mu_0^X}(h^*_{\varepsilon',\delta',Fair_0}(X) \neq h^*_{\varepsilon,\delta,0}(X)) + \mathbb{P}_{\mu_1^X}(h^*_{\varepsilon',\delta',Fair_1}(X) \neq h^*_{\varepsilon,\delta,1}(X)) \\
&\leq_\eta err^*(h^*_{\varepsilon,\delta,0}, h^*_{\varepsilon,\delta,1}) + \frac{5}{2}\Big(\frac{\log(4/\eta)}{n\theta\varepsilon_0} + \frac{\log(4/\eta)}{n\bar{\theta}\varepsilon_1} + \sqrt{\log(\frac{8}{\eta})\frac{1}{2n\theta}} + \sqrt{\log(\frac{8}{\eta})\frac{1}{2n\bar{\theta}}}\Big).
\end{aligned}$$

The privacy guarantee of $h^*_{\varepsilon',\delta',\text{Fair}}$ has two components: the privacy guarantees of the input classifiers and the Laplace mechanism employed to privately estimate $\bar{\alpha}$ and $\bar{\beta}$ (line 2 in Algorithm 2). Thus, the privacy guarantee in the above theorem follows directly from a composition result (e.g., basic composition Dwork *et al.* (2014)). However, the analysis pertaining to fairness and utility guarantees is rather long and thus deferred to the appendix. The fairness guarantee demonstrates, as expected, that perfect statistical parity can no longer be achievable when requiring privacy as well. Nevertheless, the theorem shows that $h^*_{\varepsilon',\delta',\text{Fair}}$ satisfies γ-statistical parity where $\gamma > 0$ is a small constant provided that the dataset is not highly imbalanced (i.e., θ close to 0 or 1) and n is sufficiently large (compared to $1/\varepsilon_0$ and $1/\varepsilon_1$). We remark that these assumptions were implicitly made in Zhao and Gordon (2022), in which

they ignored the error in estimating the proportion of label one in each subgroup. Finally, the utility guarantee presented in the theorem indicates that the necessary perturbations in the final prediction by $h^*_{\varepsilon',\delta',\text{Fair}}$ is almost identical to what is expected by the optimal classifier having the same level of statistical parity.

Rather than aiming to achieve a small statistical parity gap and a small utility gap with high probability, we could alternatively focus on the *average* guarantees for fairness and utility, as expounded by the next result.

Proposition 9. *The classifier $h^*_{\varepsilon',\delta',Fair} : \mathcal{X} \times \{0,1\} \to \{0,1\}$ constructed by Algorithm 2 satisfies the following two properties:*

- *We have: (Fairness guarantee)*

$$\mathbb{E}\big[\Delta_{SP}(h^*_{\varepsilon',\delta',Fair})\big] \leq \frac{1}{n\theta\varepsilon_0} + \frac{1}{n\bar{\theta}\varepsilon_1} + \sqrt{\frac{1}{4n\theta}} + \sqrt{\frac{1}{4n\bar{\theta}}},$$

- *(Utility guarantee) We have:*

$$\begin{aligned}\mathbb{E}\Big[\mathbb{P}_{\mu_0^X}(h^*_{\varepsilon',\delta',Fair_0}(X) \neq h^*_{\varepsilon,\delta,0}(X)) + \mathbb{P}_{\mu_1^X}(h^*_{\varepsilon',\delta',Fair_1}(X) \neq h^*_{\varepsilon,\delta,1}(X))\Big] \\ \leq \mathbb{E}\big[err^*(h^*_{\varepsilon,\delta,0}, h^*_{\varepsilon,\delta,1})\big] + \frac{5}{2}\Big(\frac{1}{n\theta\varepsilon_0} + \frac{1}{n\bar{\theta}\varepsilon_1} + \sqrt{\frac{1}{4n\theta}} + \sqrt{\frac{1}{4n\bar{\theta}}}\Big),\end{aligned}$$

  where $err^*(\cdot,\cdot)$ was defined in (3.4.1).

3.5 Experiments

In this section, we seek to empirically compare Algorithm 2 with the current state-of-the-art differentially private fair classifier, namely, DP-FERMI (Lowy *et al.*, 2023).

To this goal, we focus on two benchmark datasets from the UCI machine learning repository (Lichman, 2013): the Adult and Credit Card datasets, both with binary sensitive attributes and target labels.

In our experiments, we assessed two privacy configurations: $(\varepsilon' = 3, \delta' = 10^{-5})$ and $(\varepsilon' = 9, \delta' = 10^{-5})$. When applying Algorithm 2 to the Adult and Credit Card datasets, we set $\varepsilon_0 = \varepsilon_1 = 0.05$ and $\varepsilon_0 = \varepsilon_1 = 0.1$, respectively. We then feed Algorithm 2 with the decoupled classifiers, $h^*_{\varepsilon,\delta,0}$ and $h^*_{\varepsilon,\delta,1}$ with parameters (ε, δ) chosen in a way as to satisfy $\varepsilon' = \varepsilon + \varepsilon_0 + \varepsilon_1$ and $\delta' = \delta$ for each specified (ε', δ') pair. We trained these classifiers via DP-SGD. The training was conducted using Opacus (Yousefpour *et al.*, 2021), an open-source PyTorch library designed for training deep learning models with differential privacy. Accordingly, we chose the standard deviations of Gaussian noise in DP-SGD to be 3.13 and 1.5 for the Adult dataset, and to 4.44 and 2.49 for the Credit Card dataset, in order to achieve their respective privacy parameters $(\varepsilon' = 3, \delta' = 10^{-5})$ and $(\varepsilon' = 9, \delta' = 10^{-5})$. Across all experiments of DP-SGD, we consistently applied a clipping constant of 1.5 and a learning rate of 0.01. To achieve the most precise computation of privacy parameters, we utilized the PRV accountant (Gopi *et al.*, 2021), the state-of-the-art accounting method, to determine the values for (ε, δ) for the classifiers $h^*_{\varepsilon,\delta,0}$ and $h^*_{\varepsilon,\delta,1}$.

For comparability with DP-FERMI, we employed logistic regression models. DP-FERMI utilizes a loss function with a regularization constant λ to limit statistical parity violations. A higher λ imposes stricter penalties for fairness violations, often at the cost of reduced accuracy. In this framework, the desired ε' and δ' are set and the required noise levels to meet these privacy guarantees are computed. It is important to remark that DP-FERMI implicitly assumes that both the fraction of data points

in the minority subgroup and the size of the entire dataset are publicly available —similar to the assumption made in our work. All models, including Algorithm 2 and DP-FERMI, were trained for 50 epochs, except the Credit Card models using $(\varepsilon' = 9, \delta' = 10^{-5})$ parameters, which were trained for 100 epochs. Within DP-FERMI, learning rates were set at $\eta_\theta = 0.005$ and $\eta_W = 0.01$. We chose these parameters mainly because they were empirically shown in Lowy *et al.* (2023) to be optimal for the datasets under consideration. A uniform batch size of 1024 was maintained for all experiments.

While our theoretical framework focuses on the sum of prediction changes across subgroup distributions, for comparison purposes with DP-FERMI, we used overall accuracy as a utility metric. The final results are presented in Tables 3.1, 3.2, 3.3, and 3.4. All of the results are averages over 10 trials. We remark that DP-FERMI offers two privacy options: one for sensitive attributes and another for all features. We used noise parameters for all-feature privacy in DP-FERMI to compare fairly with Algorithm 2.

Table 3.1: Adult dataset $(\varepsilon' = 3, \delta' = 10^{-5})$ - PRV accountant

Method	Accuracy	Statistical Parity Gap
Algorithm 2	**0.7763**	**0.0074**
DP-FERMI ($\lambda = 0.5$)	0.7998	0.1020
DP-FERMI ($\lambda = 1$)	0.7859	0.0462
DP-FERMI ($\lambda = 1.5$)	0.7822	0.0267
DP-FERMI ($\lambda = 1.8$)	**0.7770**	**0.0182**
DP-FERMI ($\lambda = 2.5$)	0.7673	0.0099

In the Adult dataset, as shown in Table 3.1, Algorithm 2 achieves accuracy 0.7763. Setting λ at 1.8, DP-FERMI achieves a comparable accuracy (0.7770) but exhibits a statistical parity gap more than twice as large as what is guaranteed by Algorithm 2.

Table 3.2: Adult dataset ($\varepsilon' = 9, \delta' = 10^{-5}$) - PRV accountant

Method	Accuracy	Statistical Parity Gap
Algorithm 2	**0.7790**	**0.0091**
DP-FERMI ($\lambda = 0.5$)	0.8091	0.0944
DP-FERMI ($\lambda = 1$)	0.7923	0.0413
DP-FERMI ($\lambda = 1.5$)	0.7810	0.0152
DP-FERMI ($\lambda = 1.7$)	**0.7782**	**0.0121**
DP-FERMI ($\lambda = 2.5$)	0.7693	0.0030

Table 3.3: Credit Card dataset ($\varepsilon' = 3, \delta' = 10^{-5}$) - PRV accountant

Method	Accuracy	Statistical Parity Gap
Algorithm 2	**0.7844**	**0.0086**
DP-FERMI ($\lambda = 0.1$)	0.7899	0.0212
DP-FERMI ($\lambda = 0.2$)	**0.7846**	**0.0193**
DP-FERMI ($\lambda = 0.5$)	0.7777	0.0185
DP-FERMI ($\lambda = 1$)	0.7759	0.0105
DP-FERMI ($\lambda = 2.5$)	0.7669	0.0110

Table 3.4: Credit Card dataset ($\varepsilon' = 9, \delta' = 10^{-5}$) - PRV accountant

Method	Accuracy	Statistical Parity Gap
Algorithm 2	**0.7900**	**0.0056**
DP-FERMI ($\lambda = 0.25$)	0.7996	0.0188
DP-FERMI ($\lambda = 0.3$)	0.7971	0.0182
DP-FERMI ($\lambda = 0.5$)	0.7912	0.0172
DP-FERMI ($\lambda = 1$)	**0.7895**	**0.0105**
DP-FERMI ($\lambda = 2.5$)	0.7884	0.0066

Table 3.2 illustrates that Algorithm 2 and DP-FERMI with $\lambda = 1.7$ achieve similar accuracy, yet Algorithm 2 exhibits a smaller statistical parity gap (0.0091) compared to DP-FERMI (0.0121). Altogether, these tables empirically underscore that applying Algorithm 2 to the Adult dataset results in a better fairness guarantee while maintaining similar privacy and accuracy guarantees.

Similar trends are observed in the Credit Card dataset. Table 3.3 displays Algorithm 2 achieving an accuracy of 0.7844, closely matched by DP-FERMI with an accuracy of 0.7846 at $\lambda = 0.2$. However, the statistical parity gap of Algorithm 2 is less than half that of DP-FERMI (0.0086 compared to 0.0193). As shown in Table 3.4, Algorithm 2 and DP-FERMI with $\lambda = 1$ attain similar accuracy, yet Algorithm 2 exhibits a statistical parity gap of 0.0056, nearly half of DP-FERMI's 0.0105.

Chapter 4

Local DP and Pre-Processing Techniques for Fairness: Related Work

We categorize the related work into two main groups. The first group includes studies exploring the intersection of LDP and fairness. The second group consists of papers focused on fairness intervention methods for classification. Among these, studies on pre-processing methods aimed at improving fairness are the most relevant to our work.

4.1 Effect of LDP on Fairness

Several studies have examined the impact of LDP on fairness in classification. Arcolezi *et al.* (2023) empirically shows that applying LDP to sensitive attributes improves fairness with minimal utility loss compared to non-private data, highlighting

that GRR and subset selection (SS) (Wang *et al.*, 2016) offer the best accuracy-fairness-privacy trade-offs among state-of-the-art LDP mechanisms. Expanding on this, Makhlouf *et al.* (2024a) experimentally demonstrates that stronger privacy (i.e., lower ε) further enhances fairness, with greater reductions in disparity when LDP is applied to multiple sensitive attributes. They subsequently delineate this observation by characterizing conditions under which RR implies better unfairness for binary sensitive attributes (Makhlouf *et al.*, 2024b). The work most closely related to ours is Makhlouf *et al.* (2024b), as they also theoretically analyze the impact of LDP on fairness. However, their study is limited to binary sensitive attributes only. Other works focus on learning frameworks with privatized sensitive attributes. For instance, Mozannar *et al.* (2020) adapts non-discriminatory classifiers to work with privatized attributes using GRR, offering theoretical performance guarantees on utility and fairness. Similarly, Chen *et al.* (2022) considers the scenario of semi-private sensitive attributes.

4.2 Fairness Intervention Methods

Fairness intervention methods are applied at different stages to mitigate bias: pre-processing before training (Calders *et al.*, 2009; Wang *et al.*, 2019; Kamiran and Calders, 2012; Celis *et al.*, 2020; Calmon *et al.*, 2017; Hajian and Domingo-Ferrer, 2012; Chakraborty *et al.*, 2021; Peng *et al.*, 2022; Gohar *et al.*, 2023; Madras *et al.*, 2018; Zemel *et al.*, 2013), in-processing during model training (Lowy *et al.*, 2021; Cho *et al.*, 2020a,b; Jiang *et al.*, 2020; Mary *et al.*, 2019; Prost *et al.*, 2019; Zhang *et al.*, 2018; Agarwal *et al.*, 2018; Zafar *et al.*, 2017), and post-processing after generating predictions (Wei *et al.*, 2020, 2021; Chzhen *et al.*, 2019; Pleiss *et al.*, 2017; Kim

et al., 2020; Jiang and Nachum, 2020; Yang *et al.*, 2020; Alghamdi *et al.*, 2022). Among these, pre-processing methods offer the most flexibility within the data science pipeline, as they operate independently of the modeling algorithm (Calmon *et al.*, 2017).

4.3 Pre-Processing Techniques for Fairness

Pre-processing techniques for bias mitigation in fairness literature involve making changes to training data. Some methods modify values within the training data, such as altering ground truth labels (relabeling) (Calders *et al.*, 2009; Hajian and Domingo-Ferrer, 2012; Kamiran and Calders, 2012) or adjusting other features (perturbation) (Wang *et al.*, 2019). Kamiran and Calders (2012) reduces bias by re-weighting existing data points, while Wang *et al.* (2019) perturbs the input distribution for disadvantaged groups to create a counterfactual distribution, particularly targeting binary sensitive attributes.

Another category of work considers sampling methods. Sampling methods adjust training data by changing sample distributions (e.g., adding or removing samples) or adapting their influence on training (Chakraborty *et al.*, 2021; Celis *et al.*, 2020). Celis *et al.* (2020) proposes an optimization-based framework to learn distributions over the data domain that stay close to the empirical distribution. This technique is applicable to datasets with discrete or categorical attributes, both sensitive and non-sensitive. Other methods augment training data with additional, ideally unbiased features (Madras *et al.*, 2018). Finally, some techniques learn transformations of the training data to reduce bias while retaining as much information as possible

(Zemel *et al.*, 2013; Calmon *et al.*, 2017). Calmon *et al.* (2017) introduces an optimization algorithm that modifies non-sensitive features and labels while keeping sensitive attributes unchanged, focusing on datasets with categorical or discrete attributes. Our method aligns with perturbation-based approaches, as we perturb the sensitive attribute using LDP.

4.4 Overview of Our Approach

Our approach aligns closely with pre-processing methods, employing LDP as a data pre-processing technique aimed at enhancing fairness. Building on the recent findings about the relationship between LDP and fairness (Arcolezi *et al.*, 2023; Makhlouf *et al.*, 2024b,a), we investigate the use of LDP as a pre-processing technique to reduce unfairness. Our approach formulates the problem of identifying the optimal LDP-based pre-processing mechanism to minimize data unfairness. We also offer a theoretical analysis that explains how this mechanism influences classification fairness. Finally, we validate the effectiveness of this pre-processing mechanism through extensive experiments across various datasets and fairness metrics, demonstrating its superior performance in reducing classification unfairness.

Chapter 5

Local DP as a Pre-processing Technique for Fairness in Classification: Main Results

5.1 Notation

In this section, we consider a binary classification setting with a joint distribution P_{XAY} over the triplet $T = (X, A, Y)$, where $X \in \mathcal{X} \subset \mathbb{R}^d$ represents the feature vector of non-sensitive attributes. The sensitive attribute A may vary depending on the scenario: in the non-binary case, $A \in \{1, 2, \ldots, k\}$ (with $k \geq 2$), while in the binary case, $A \in \{0, 1\}$. The target output $Y \in \{0, 1\}$ is the label to be predicted. We introduce Z, the perturbed version of the sensitive attribute A, which results from applying an LDP mechanism. When such a mechanism is applied to the sensitive attribute in a dataset, the triplet (X, A, Y) becomes (X, Z, Y), where Z is a random variable derived from perturbing A. For the distribution P_{XAY}, or a dataset D with

data points independently sampled from P_{XAY}, we define $p_i = \Pr(A = i)$, representing the probability that $A = i$ for $i \in \{1, 2, \ldots, k\}$, and $p_{1|i} = \Pr(Y = 1 \mid A = i)$. While in practice we typically work with datasets and the true data distribution is unknown, we treat the dataset D and the distribution P_{XAY} interchangeably in this work. All metrics can be defined in a similar way for both the dataset and the distribution. The set $\{1, 2, \ldots, k\}$ is denoted as $[k]$, and the symbol $\lfloor \cdot \rceil$ is used to indicate rounding to the nearest positive integer.

5.2 Preliminaries

In this section, we present the key definitions used throughout the chapter and provide an overview of commonly used LDP mechanisms.

5.2.1 Data Unfairness Metrics

Definition 10 (Data unfairness Δ (Calmon *et al.*, 2017))**.** *Let $X \in \mathcal{X}$, $A \in [k]$, and $Y \in \{0, 1\}$ be random variables representing non-sensitive features, sensitive attributes, and labels, with a joint distribution P_{XAY}. Each data point in the dataset D is a triplet (x_i, a_i, y_i), where $(x_i, a_i, y_i) \sim P_{XAY}$. The data unfairness metric $\Delta(D)$ associated with D or P_{XAY} is defined as:*

$$\Delta(D) := \max_{a \in [k]} \left| \frac{\Pr(Y = 1 \mid A = a)}{\Pr(Y = 1)} - 1 \right|.$$

Definition 11 (Data unfairness Δ' (Kamiran and Calders, 2012))**.** *Let $X \in \mathcal{X}$, $A \in [k]$, and $Y \in \{0, 1\}$ be random variables representing non-sensitive features, sensitive attributes, and labels, with a joint distribution P_{XAY}. Each data point in*

the dataset D is a triplet (x_i, a_i, y_i), where $(x_i, a_i, y_i) \sim P_{XAY}$. The data unfairness metric $\Delta'(D)$ associated with D or P_{XAY} is defined as:

$$\Delta'(D) := \max_{a,a' \in [k]} \Big| \Pr(Y = 1 \mid A = a) - \Pr(Y = 1 \mid A = a') \Big|.$$

Both data unfairness metrics, Δ and Δ', capture the dependence (or independence) of the actual labels on the sensitive attribute. It can be shown that $\Delta(D) \leq c_1 \Delta'(D)$ and $\Delta'(D) \leq c_2 \Delta(D)$ for some constants c_1 and c_2 that depend on the marginal distribution P_Y of the joint distribution P_{XAY} (see Lemma 16). Therefore, the metrics Δ and Δ' are essentially the same. However, for technical reasons that will become clear in the following sections, we use Definition 11 to formulate the optimal LDP-based pre-processing problem for binary sensitive attributes in Section 5.4 and to establish the relationship between data unfairness and classification unfairness in Section 5.6. In Section 5.5, we utilize Definition 10 to formulate the optimal LDP-based pre-processing problem for non-binary sensitive attributes.

5.2.2 Discrimination-Accuracy Optimality

Definition 12 (Discrimination-accuracy optimal classifier (Kamiran and Calders, 2012))**.** *Let h and h' be two classifiers. We say that h dominates h' if the accuracy of h is greater than or equal to that of h', and the discrimination of h (measured with respect to an unfairness metric such as equalized opportunity or statistical parity) is at most that of h'. Classifier h strictly dominates h' if at least one of these inequalities is strict. Given a set of classifiers $\mathcal{H}$, we call a classifier $h \in \mathcal{H}$ optimal with respect to discrimination and accuracy (DA-optimal) in $\mathcal{H}$ if no other classifier in $\mathcal{H}$ strictly dominates h.*

5.2.3 Common LDP Mechanisms

Generalized randomized response (GRR)(Warner, 1965; Kairouz *et al.*, 2014): Given a sensitive attribute A taking values in $[k]$, the generalized randomized response (GRR) mechanism perturbs the true value $a \in [k]$ to preserve privacy. GRR outputs the true value a with probability π, and any other possible value $a' \in [k] \setminus \{a\}$ with a different probability $\bar{\pi}$. Specifically, for the perturbed output Z:

$$\forall z \in [k]: \quad \Pr(Z = z \mid A = a) = \begin{cases} \pi = \frac{e^{\varepsilon}}{e^{\varepsilon}+k-1} & \text{if } z = a, \\ \bar{\pi} = \frac{1}{e^{\varepsilon}+k-1} & \text{if } z \neq a, \end{cases}$$

In binary A case, the mechanism is known as randomized response (RR).

Subset selection (SS) (Wang *et al.*, 2016): Given a sensitive attribute A taking values in $[k]$, SS mechanism perturbs the true value $a \in [k]$ by reporting a subset of values $\Omega \subseteq [k]$. The mechanism aims to include the true value a in the subset Ω with a higher probability than any other value in $[k] \setminus \{a\}$. The optimal subset size that minimizes variance is $\omega = \left\lfloor \frac{k}{e^{\varepsilon}+1} \right\rceil$. The SS mechanism proceeds as follows:

1. Initialize an empty subset Ω.

2. Add the true value a to Ω with probability

$$p = \frac{\omega e^{\varepsilon}}{\omega e^{\varepsilon} + k - \omega}.$$

3. Fill Ω as follows:

   - If $a \in \Omega$, sample $\omega - 1$ additional values uniformly at random (without replacement) from $[k] \setminus \{a\}$ and add them to Ω.

- If $a \notin \Omega$, sample ω values uniformly at random (without replacement) from $[k] \setminus \{a\}$ and add them to Ω.

Finally, the user sends the subset Ω to the server as the perturbed value. As we can see, when $\omega = 1$, SS is equivalent to GRR.

5.3 GRR and Data Unfairness

Let $X \in \mathcal{X}$, $A \in [k]$, and $Y \in \{0, 1\}$ be random variables representing non-sensitive features, sensitive attributes, and labels, respectively, with a joint distribution P_{XAY}. Each data point in the dataset D is a triplet (x_i, a_i, y_i), where $(x_i, a_i, y_i) \sim P_{XAY}$. Now, suppose we independently perturb the sensitive attribute a_i of each data point in D using GRR to generate a new dataset D^{ε}_{GRR}, keeping the same x_i and y_i. The resulting dataset is $D^{\varepsilon}_{GRR} = \{(x_i, z_i, y_i)\}_{i=1}^{n}$, where z_i is the noisy version of a_i obtained via GRR. We define the data unfairness $\Delta'(D^{\varepsilon}_{GRR})$ in a similar manner to Definition 11 as:

$$\Delta'(D^{\varepsilon}_{GRR}) := \max_{z,z' \in [k]} \Big| \Pr(Y = 1 \mid Z = z) - \Pr(Y = 1 \mid Z = z') \Big|,$$

where Z is the random variable representing the perturbed sensitive attribute after applying GRR to A. These probabilities depend on both the original joint distribution P_{XAY} and the randomness introduced by GRR.

In the following lemma, we analyze the impact of applying GRR to the sensitive attributes of a dataset on data unfairness.

Lemma 13. *For binary Y and non-binary A, applying GRR to the sensitive attributes A of a dataset results in $\Delta'(D^{\varepsilon}_{GRR}) \leq \Delta'(D)$.*

This implies that applying GRR to the sensitive attributes of a dataset can help reduce data unfairness. A natural question that follows is: if GRR can reduce data unfairness, can we identify the optimal ε-LDP mechanism for minimizing data unfairness? In Section 5.4, we address this by formulating the optimal LDP mechanism for enhancing data fairness in the case of binary sensitive attributes.

5.4 Optimal LDP Mechanism with Binary Sensitive Attribute

Observing that GRR, as one example of an LDP mechanism, can reduce data unfairness, we now turn to the question of whether it is possible to design an optimal LDP mechanism specifically aimed at minimizing data unfairness. To begin, we consider the case of a binary sensitive attribute for simplicity.

Consider a dataset $D = \{(x_i, a_i, y_i)\}_{i=1}^n$, where $a_i \in \{0, 1\}$ represents the sensitive attribute for each data point, and y_i is the corresponding label. We independently perturb the sensitive attribute a_i for each data point using an ε-LDP mechanism M, resulting in a new dataset D_M^ε. This new dataset retains the original features x_i and labels y_i, but with perturbed sensitive attributes z_i. Thus, the transformed dataset is $D_M^\varepsilon = \{(x_i, z_i, y_i)\}_{i=1}^n$, where z_i is the noisy version of a_i generated by the mechanism M, which is a randomized mapping characterized as follows:

$$\Pr(Z = z \,|\, A = 0) = \begin{cases} p & \text{if } z = 0 \\ 1 - p & \text{if } z = 1 \end{cases}$$

$$\Pr(Z = z \mid A = 1) = \begin{cases} q & \text{if } z = 1 \\ 1 - q & \text{if } z = 0 \end{cases}, \tag{5.4.1}$$

where Z is the random variable resulting from applying the mechanism M to A. In this mechanism, p and q represent the probabilities of correctly reporting the sensitive attribute when the original values are 0 and 1, respectively, and these probabilities may differ. RR is a specific case of this mechanism where $p = q = \frac{e^\varepsilon}{e^\varepsilon+1}$.

For a given ε, the goal is to find the optimal ε-LDP mechanism. Specifically, we seek a mechanism, denoted by M, such that when applied to the original data, the resulting unfairness metric $\Delta'(D^\varepsilon_M)$ is minimized in comparison to $\Delta'(D)$. In essence, the aim is to minimize the ratio $\frac{\Delta'(D^\varepsilon_M)}{\Delta'(D)}$. Note that $\Delta'(D)$ is constant since it depends only on the distribution of the original data (P_{XAY}) and not on the parameters of the LDP mechanism.

If we consider the objective function $\min\limits_{\varepsilon-\text{LDP}M} \frac{\Delta'(D^\varepsilon_M)}{\Delta'(D)}$, the optimal mechanism for fairness would be a completely random mechanism, where the probability of misreporting sensitive attributes is $1 - p = 1 - q = \frac{1}{2}$. A fully random mechanism, with $p = q = \frac{1}{2}$, satisfies ε-LDP for any $\varepsilon \geq 0$. However, to achieve a more meaningful balance between privacy and fairness, we propose the following refined objective function:

$$\min_{\substack{\varepsilon_0-\text{LDP}\, M \\ \varepsilon_0 \geq \varepsilon}} \frac{\Delta'(D^{\varepsilon_0}_M)}{\Delta'(D)}. \tag{5.4.2}$$

This approach considers all ε_0-LDP mechanisms where ε_0 is not more private than ε (i.e., $\varepsilon_0 \geq \varepsilon$). This ensures that for a given ε, the mechanism does not compromise utility by using smaller privacy parameters, and only mechanisms with non-trivial

utility are considered. For a given ε, the goal is to find the optimal ε-LDP mechanism, specifically the optimal parameters p^* and q^*, such that the objective function (5.4.2) is minimized. The following theorem presents the optimal ε-LDP mechanism for the case of a binary sensitive attribute.

Theorem 14. *Consider the case of binary Y and binary A, where $p_{1|0} \leq p_{1|1}$. Let (p^*, q^*) represent the optimal parameters that minimize the objective function defined in (5.4.2). The optimal LDP mechanism is determined as follows:*

$$\textit{If} \quad p_0 < p_1, \quad (p^*, q^*) = \left(1 - \frac{e^{-\varepsilon}}{2}, \frac{1}{2}\right),$$

$$\textit{If} \quad p_1 < p_0, \quad (p^*, q^*) = \left(\frac{1}{2}, 1 - \frac{e^{-\varepsilon}}{2}\right),$$

where p and q are the parameters of the general LDP mechanism as defined in (5.4.1).

This theorem identifies the optimal mechanism for applying LDP as a pre-processing strategy in the binary sensitive attribute case. Essentially, LDP is used as a tool to pre-process data and minimize data unfairness. However, sensitive attributes are not always binary, leading to the natural inquiry of how to determine the optimal pre-processing mechanism under LDP constraints for non-binary sensitive attributes. In the next section, we explore this issue.

5.5 Optimal LDP Mechanism with Non-Binary Sensitive Attribute

As demonstrated in Theorem 14, the optimal pre-processing mechanism under LDP for a binary sensitive attribute can be expressed in closed form, i.e., the optimal solution (p^*, q^*) can be computed based on data distribution and the parameter ε. In this section, we extend this result to the case of non-binary sensitive attributes. We begin by defining the problem setting for non-binary sensitive attributes and then discuss how the optimal mechanism can be derived under this new setting.

We want to find the optimal LDP mechanism for a binary Y and non-binary A case. Let $A \in [k]$, and let $\mathbf{Q}$ be a $k \times k$ matrix containing the parameters of the LDP mechanism. The randomized mechanism changes the sensitive attribute of the original data from $A = a$ to $Z = z$ using the parameters of $\mathbf{Q}$, defined as:

$$\Pr(Z = j \mid A = i) = q_{ij} \qquad \forall i, j \in [k],$$

where q_{ij} is the element in the i-th row and j-th column of $\mathbf{Q}$. The matrix $\mathbf{Q}$ should satisfy the following constraints:

1. **LDP Constraint:** The matrix $\mathbf{Q}$ should satisfy ε-LDP, i.e.,

$$q_{ij} - e^{\varepsilon} q_{i'j} \leq 0 \qquad \forall i, i', j \in [k] \quad \text{with } i \neq i'. \tag{5.5.1}$$

2. **Row-Stochastic Constraint:** The matrix $\mathbf{Q}$ should be row-stochastic, i.e.,

$$\sum_{j=1}^{k} q_{ij} = 1 \qquad \forall i \in [k], \tag{5.5.2}$$

$$q_{ij} \geq 0 \qquad \forall i, j \in [k]. \tag{5.5.3}$$

3. **Truthfulness Constraint:** The probability of truly representing the sensitive attribute should be larger than or equal to the probability of misrepresenting it, i.e.,

$$q_{ii} \geq q_{ij} \qquad \forall i, j \in [k], \tag{5.5.4}$$

$$q_{jj} \geq q_{ij} \qquad \forall i, j \in [k]. \tag{5.5.5}$$

4. **Utility Constraint:** The mechanism should satisfy a utility metric, where the probability of error (i.e., total probability of $Z \neq A$) should be smaller than some predefined constant ζ. This constraint can be formulated as:

$$\Pr(A = Z) \geq 1 - \zeta.$$

We know:

$$\Pr(A = Z) = \sum_{i=1}^{k} \Pr(Z = i, A = i) = \sum_{i=1}^{k} \Pr(Z = i \mid A = i) \Pr(A = i) = \sum_{i=1}^{k} q_{ii} p_i.$$

Therefore, the utility constraint becomes:

$$\sum_{i=1}^{k} q_{ii} p_i \geq 1 - \zeta \tag{5.5.6}$$

This utility constraint aims to regulate the trade-off between fairness and utility. Without the utility constraint, if we only consider the data fairness objective function,

we may converge to a scenario where perfect data fairness is achieved at the cost of poor utility. This constraint prevents such a scenario. The parameter ζ allows us to control the utility of the LDP mechanism.

Given (5.5.5), (5.5.1) can be reduced to:

$$q_{jj} - e^{\varepsilon} q_{ij} \leq 0 \qquad \forall i, j \in [k]. \tag{5.5.7}$$

Altogether, the optimization problem is as follows:

$$\begin{aligned} \min_{\mathbf{Q}} \quad & \Delta(D_M^{\varepsilon}) && && (5.5.8)\\ \text{s.t.} \quad & q_{jj} - e^{\varepsilon} q_{ij} \leq 0 && && \forall i, j \in [k] \\ & \sum_{j=1}^{k} q_{ij} = 1 && && \forall i \in [k] \\ & q_{ij} \geq 0, \quad q_{ii} \geq q_{ij} \quad q_{jj} \geq q_{ij}, && && \forall i, j \in [k] \\ & \sum_{i=1}^{k} q_{ii} p_i \geq 1 - \zeta \end{aligned}$$

Now, we need to write $\Delta(D_M^{\varepsilon})$ in terms of the parameters of the optimization problem, i.e., entries of matrix $\mathbf{Q}$. It can be shown that:

$$\Delta(D_M^{\varepsilon}) = \max_{a \in [k]} \left| \frac{\Pr(Y = 1 \mid Z = a)}{\Pr(Y = 1)} - 1 \right| = \max_{a \in [k]} \left| \frac{\sum_{j=1}^{k} p_{1|j} p_j q_{ja}}{\sum_{j=1}^{k} \Pr(Y = 1) p_j q_{ja}} - 1 \right|.$$

To transform the equality constraint into an inequality constraint, we assume that $q_{ii} = 1 - \sum_{\substack{j=1 \\ j \neq i}}^{k} q_{ij} \quad \forall i \in [k]$. With this assumption, optimization problem (5.5.8) can

be reformulated as:

$$\min_{\mathbf{Q}} \max_{a\in[k]} \left| \frac{\sum_{\substack{j=1\\ j\neq a}}^{k} p_{1|j}p_j q_{ja} + p_{1|a}p_a q_{aa} - \sum_{\substack{j=1\\ j\neq a}}^{k} \Pr(Y=1)p_j q_{ja} - \Pr(Y=1)p_a q_{aa}}{\sum_{\substack{j=1\\ j\neq a}}^{k} \Pr(Y=1)p_j q_{ja} + \Pr(Y=1)p_a q_{aa}} \right| \tag{5.5.9}$$

$$\text{s.t.} \quad \Big(1 - \sum_{\substack{a=1\\ a\neq j}}^{k} q_{ja}\Big) - e^{\varepsilon} q_{ij} \leq 0 \qquad \forall i,j \in [k], \quad i \neq j$$

$$\sum_{\substack{a=1\\ a\neq i}}^{k} q_{ia} \leq 1 \qquad \forall i \in [k]$$

$$q_{ij} \geq 0, \quad 1 - \sum_{\substack{a=1\\ a\neq i}}^{k} q_{ia} \geq q_{ij}, \quad 1 - \sum_{\substack{a=1\\ a\neq j}}^{k} q_{ja} \geq q_{ij} \qquad \forall i,j \in [k], \quad i \neq j$$

$$\sum_{i=1}^{k} \Big(1 - \sum_{\substack{a=1\\ a\neq i}}^{k} q_{ia}\Big) p_i \geq 1 - \zeta$$

Note that the parameters of this optimization problem are the non-diagonal entries of the matrix $\mathbf{Q}$, rather than all of its entries. This is because we know that $q_{aa} = 1 - \sum_{\substack{j=1\\ j\neq a}}^{k} q_{aj}$ for all $a \in [k]$. The optimization problem (5.5.9) can be viewed as a min-max linear fractional program, as described in Jiao and Li (2022). For fixed values of ε and ζ, this problem can be solved numerically using the branch-and-bound method, as outlined in Jiao and Li (2022). Thus, for any given data distribution, we can numerically solve the problem to determine the optimal pre-processing mechanism with the desired ε parameter for LDP and ζ error parameter. We choose Definition 10 over Definition 11 because the optimization problem (5.5.8) can be expressed as a min-max linear fractional program when using the data unfairness metric Δ.

5.6 Data Fairness Leads to Classification Fairness

In previous sections, we introduced an optimal LDP-based mechanism designed to minimize data unfairness. However, the ultimate goal extends beyond minimizing data unfairness itself. The focus is to ensure that, when a classifier is trained on such data, the resulting classification unfairness is also minimized. Since classification unfairness depends on the specific classifier being used, it is challenging to make a universal statement in this regard. In this section, we will explore a condition that, if satisfied by the classifier, allows us to claim that reducing data unfairness will lead to a corresponding reduction in classification unfairness.

Utilizing the concept of DA-optimal classifiers, we can now establish a relationship between data unfairness and classification unfairness, as outlined in the following theorem.

Theorem 15. *Let P_{XAY} and Q_{XAY} be two joint data distributions over $\mathcal{X} \times \{0,1\} \times \{0,1\}$, and let $\mathcal{H}^*$ denote the class of all classifiers satisfying $\Pr(\hat{Y}=1) = \Pr(Y=1)$. Assume that for all $a \in \{0,1\}$, the marginal distributions of the sensitive attribute are equal, i.e., $\Pr_{(X,A,Y)\sim P_{XAY}}(A=a) = \Pr_{(X,A,Y)\sim Q_{XAY}}(A=a)$ for all $a \in \{0,1\}$, and that the data unfairness under P_{XAY} is no greater than that under Q_{XAY}, i.e., $\Delta'(P_{XAY}) \leq \Delta'(Q_{XAY})$.*

If the classifiers h_P and h_Q are DA-optimal within $\mathcal{H}^*$, have equal accuracy, and are learned on data drawn from P_{XAY} and Q_{XAY} respectively, then the following inequality holds:

$$\Delta_{\mathrm{SP}}(h_P) \leq \Delta_{\mathrm{SP}}(h_Q).$$

This theorem establishes a link between data unfairness and classification unfairness using the DA-optimality condition (Kamiran and Calders, 2012). When classifiers satisfy this condition, it ensures a meaningful relationship between data and classification unfairness, justifying why reducing data unfairness can lead to models that make less discriminatory decisions. The statistical parity gap is used as a measure of classification fairness in this context. We employ Definition 11 as a data unfairness metric to establish this connection because it provides a clear relationship between the statistical parity gap and the Δ' unfairness metric under the DA-optimality condition. Both Definition 4 and Definition 11 capture the relationship between predicted or actual labels and sensitive attributes in the same way.

Having theoretically examined the optimal LDP-based pre-processing mechanisms for minimizing data unfairness, the next section will present experiments evaluating the performance of these mechanisms in classification tasks. These experiments will address both binary and non-binary sensitive attribute optimization problems.

5.7 Experiments

Optimal mechanisms discussed in Sections 5.4 and 5.5 are designed to minimize data unfairness under utility and LDP constraints. In this section, we empirically demonstrate the effectiveness of these mechanisms in classification settings. We compare our optimal mechanism (OPT) with GRR and SS. As shown in Arcolezi *et al.* (2023), GRR and SS generally provide the best privacy-utility-fairness trade-off across various datasets. We compare OPT with GRR and SS in both binary and non-binary sensitive attribute cases, evaluating both utility and fairness metrics across different datasets. In Section 5.7.1, we outline the experimental setup for this comparison,

including the datasets used and the detailed procedures of our experiments. Section 5.7.2 presents the results for the binary sensitive attribute case, while Section 5.7.3 showcases the empirical results for the non-binary sensitive attribute case. Finally, in Section 5.7.4, we compare OPT with Fair Projection (Alghamdi *et al.*, 2022), the state-of-the-art post-processing fairness intervention framework (Wang *et al.*, 2024; Denis *et al.*, 2024).

5.7.1 Experimental Setup for Comparing LDP Mechanisms

Our experimental setup follows a similar approach to Arcolezi *et al.* (2023), using the Adult (Lichman, 2013) and Law School Admissions Council (LSAC) (Wightman, 1998) datasets. In the Adult dataset, the classification label Y represents an individual's income, which is binary based on a threshold of \$26k. For the LSAC dataset, the target label Y indicates whether a candidate has passed the bar exam, which is also a binary outcome ($Y \in \{0, 1\}$). Depending on the specific experiment—whether binary or non-binary sensitive attributes are used—the sensitive attribute varies between race, gender, a combination of race and gender, or family income.

For each experimental case, we apply the LDP mechanisms GRR, SS, and OPT for different values of ε (the LDP parameter) and compare the results with the non-DP case, where no pre-processing is applied to the sensitive attributes and classification is performed directly. The classifier $\hat{h}$ is implemented using the state-of-the-art LGBM (Ke *et al.*, 2017), with an 80/20 training/testing split. All experiments are averaged over 20 different random seeds to reduce the influence of randomness.

In our experiments, individual predictions for both the non-private and private versions are obtained by applying the models to the original testing data without

LDP mechanisms. As utility metrics, we report accuracy and F1-score, while fairness metrics include the equalized opportunity gap $\Delta_{\text{EO}}(\hat{h})$ and statistical parity gap $\Delta_{\text{SP}}(\hat{h})$.

5.7.2 Comparison of LDP Mechanisms: Binary Sensitive Attribute

In this section, we apply three LDP mechanisms—OPT, GRR, and SS—to binary sensitive attributes. We then train a classifier on the perturbed datasets and compare the utility and fairness metrics of these mechanisms against a non-DP case, where no LDP pre-processing is applied to the data. In the binary case, OPT refers to the mechanism defined in Theorem 14. For both Adult and LSAC datasets, we considered gender to be a sensitive attribute.

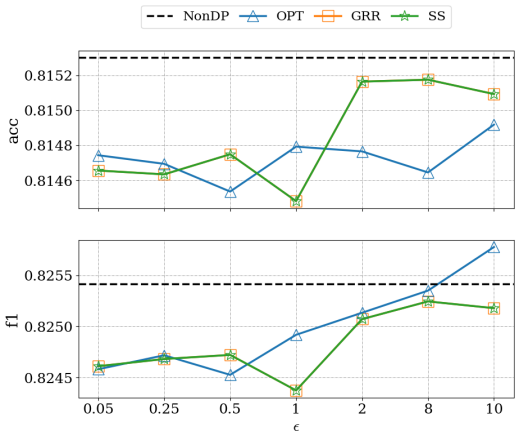

Figure 5.1: Utility metrics for Adult dataset with binary sensitive attribute `'gender'`

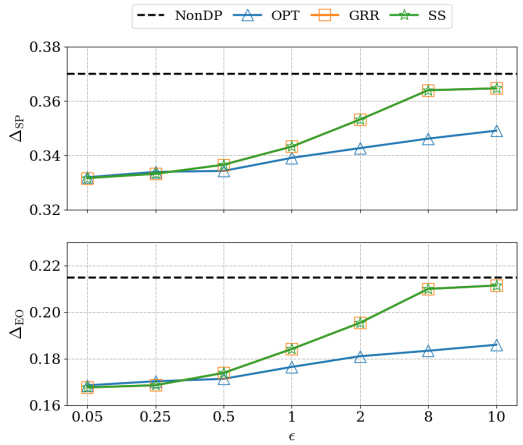

Figure 5.2: Fairness metrics for Adult dataset with binary sensitive attribute `'gender'`

As shown in Figures 5.1 and 5.3, applying OPT has little to no effect on utility metrics such as F1-score and accuracy across different ε values, compared to the

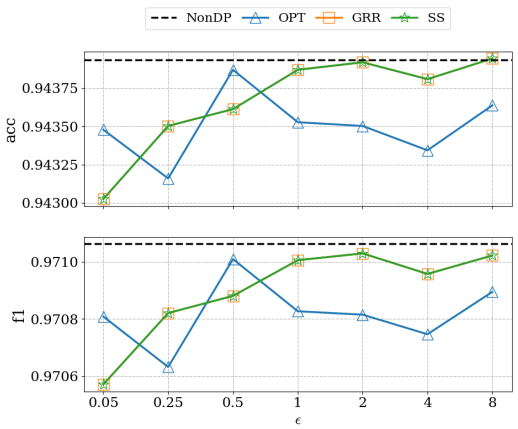

Figure 5.3: Utility metrics for LSAC dataset with binary sensitive attribute `'gender'`

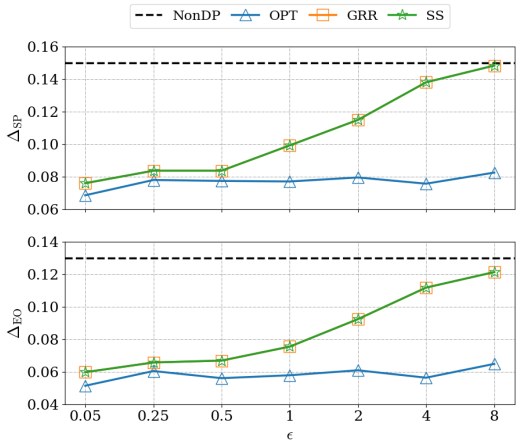

Figure 5.4: Fairness metrics for LSAC dataset with binary sensitive attribute `'gender'`

other LDP mechanisms and even the non-DP case. However, Figures 5.2 and 5.4 demonstrate consistently smaller classification unfairness gaps for OPT compared to GRR, SS, and the non-DP case.

For smaller ε values, OPT outperforms the other mechanisms in terms of classification fairness. As ε increases, the gap in unfairness between OPT and GRR/SS widens. In particular, for larger ε values, both $\Delta_{\mathrm{SP}}(\hat{h})$ and $\Delta_{\mathrm{EO}}(\hat{h})$ for OPT are almost half of those observed for GRR, SS, and the non-DP case for the LSAC dataset. For the Adult dataset, OPT shows at least a 2% reduction in both unfairness metrics for larger ε values.

Note that, in the binary case, OPT does not converge to the non-private scenario as ε increases. This is because, in the optimal mechanism, even for large ε values, one of the sensitive attributes is always perturbed with probability $\frac{1}{2}$, representing an optimal LDP-based pre-processing strategy. Also, we should note that in a binary sensitive attribute case, SS and RR are essentially the same mechanisms.

5.7.3 Comparison of LDP Mechanisms: Non-Binary Sensitive Attribute

In this section, we use the same datasets but with different sensitive attribute features to handle non-binary cases. For the Adult dataset, we consider two scenarios. In the first scenario, the sensitive attribute is race, which has five distinct values ($k = 5$). In the second scenario, we combine race and gender into a single attribute, resulting in a combined sensitive attribute with ten distinct values ($k = 10$). For the LSAC dataset, we consider family income as the non-binary sensitive attribute, also with five distinct values ($k = 5$).

For the non-binary sensitive attribute case, we formulate the min-max linear fractional program associated with finding the optimal LDP mechanism, as described in (5.5.9). We then solve this problem numerically, following the approach in Jiao and Li (2022). Given specific values of ε and ζ, we can obtain the optimal mechanism by solving the program. For the error probability ζ, we use the smallest value for which the problem (5.5.9) remains feasible.

After determining the optimal mechanism, we apply it to the sensitive attributes. Similar to the binary case, we compare the performance of the optimal mechanism with GRR, SS, and the non-DP case using both fairness and utility metrics after classification.

As shown in Figures 5.5, 5.6, 5.7, 5.8, 5.9, and 5.10, the accuracy and F1-score remain very similar across all LDP mechanisms, including OPT, GRR, and SS, consistent with the results from the binary sensitive attribute experiments. OPT performs similarly to GRR and SS for smaller values of ε in all three experiments. However, beyond a certain threshold of ε, OPT begins to outperform GRR and SS in fairness

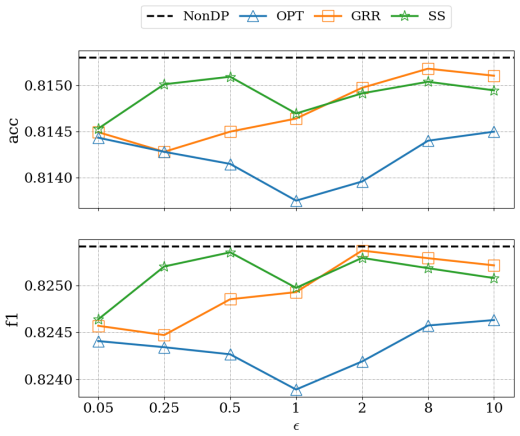

Figure 5.5: Utility metrics for Adult dataset with non-binary sensitive attribute `'race'` with $k = 5$

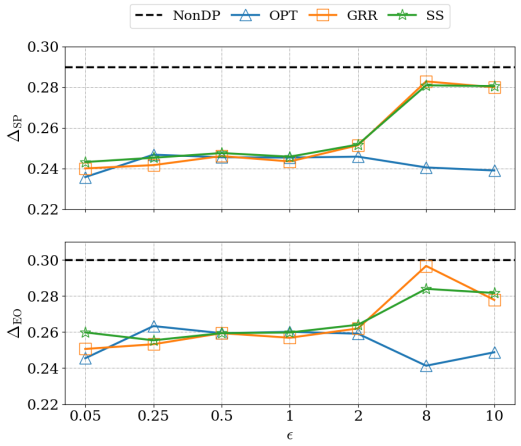

Figure 5.6: Fairness metrics for Adult dataset with non-binary sensitive attribute `'race'` with $k = 5$

metrics, including the statistical parity gap and the equalized opportunity gap, underscoring the effectiveness of OPT as an optimal LDP-based pre-processing method.

Although we do not have a closed-form solution for the OPT mechanism in a non-binary A case, the mechanism obtained by solving the optimization problem (5.5.9) proves to be an effective pre-processing method, reducing unfairness while having a negligible impact on utility metrics. It is important to note that the values of the unfairness metrics, $\Delta_{\text{SP}}(\hat{h})$ and $\Delta_{\text{EO}}(\hat{h})$, tend to increase as the support size of the sensitive attribute increases. This is because these metrics are defined as worst-case differences.

We combined the sensitive attributes race and gender to increase the support size of the sensitive attribute, allowing us to evaluate how OPT performs with larger support sizes (k). While increasing k does result in a longer computation time for solving the min-max linear fractional optimization problem (5.5.9), the approach remains effective in reducing classification unfairness.

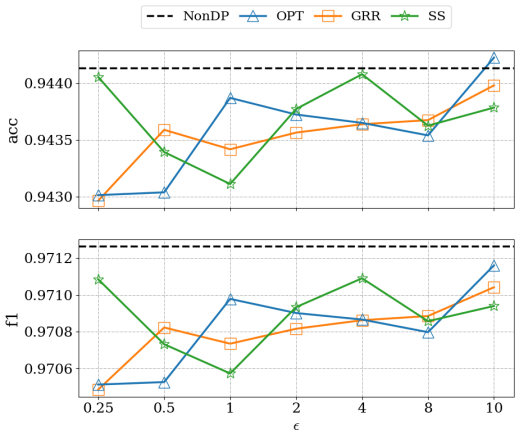

Figure 5.7: Utility metrics for LSAC dataset with non-binary sensitive attribute `'family income'` with $k = 5$

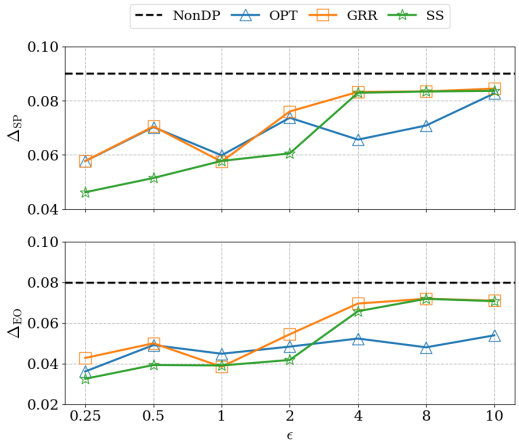

Figure 5.8: Fairness metrics for LSAC dataset with non-binary sensitive attribute `'family income'` with $k = 5$

5.7.4 Comparing Optimal LDP Mechanism with Fair Projection

In previous sections, experimental results demonstrated that the optimal LDP-based mechanism outperforms standard LDP mechanisms in classification fairness while having a minimal impact on utility. In this section, we compare OPT with Fair Projection (Alghamdi *et al.*, 2022), the state-of-the-art post-processing fairness intervention framework.

For each dataset, we apply OPT to the sensitive attribute, using three different values of ε. After this transformation, we proceed with the classification task on the pre-processed dataset. We compare OPT with a Non-Fair classification baseline, where no fairness intervention is applied, and classification is performed directly on the original dataset without any pre-processing. Additionally, we compare OPT with Fair Projection which performs classification on the original dataset, followed by

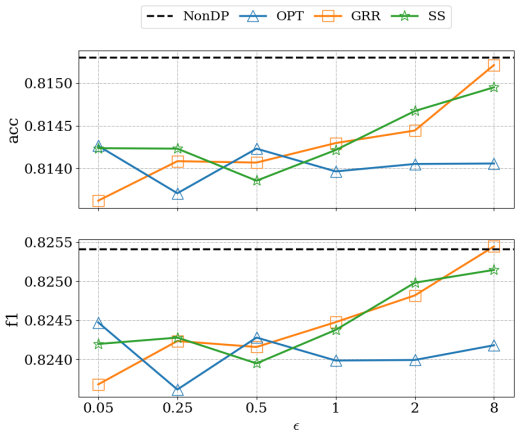

Figure 5.9: Utility metrics for Adult dataset with combined non-binary sensitive attribute `'race-gender'` with $k = 10$

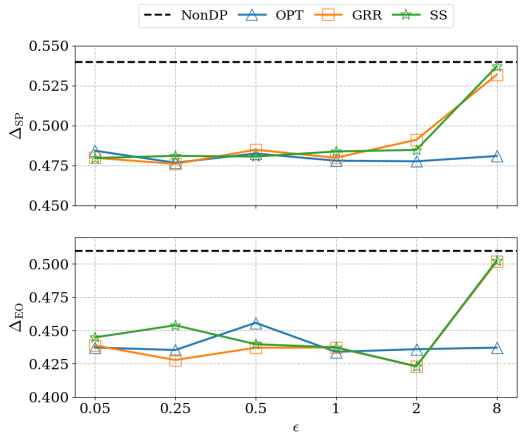

Figure 5.10: Fairness metrics for Adult dataset with combined non-binary sensitive attribute `'race-gender'` with $k = 10$

a post-processing step to obtain a final model that satisfies fairness criteria. The experiments in this section are conducted on a binary sensitive attribute. The same experimental approach can be applied to a non-binary sensitive attribute, with the optimal mechanism obtained by algorithmically solving the optimization problem (5.5.9).

Fair Projection introduces an accuracy-fairness tradeoff that allows adjustments based on the desired level of fairness in the classifications. To ensure a fair comparison with OPT, we select the point on this tradeoff curve where the unfairness metric of Fair Projection matches that of the optimal mechanism. This alignment enables a utility comparison under a fixed level of unfairness. We use cross-entropy and KL-divergence as post-processing loss functions, following the approach in Alghamdi *et al.* (2022). These represent different instantiations of Fair Projection (FairProjection-CE and FairProjection-KL). We consider two fairness constraints: mean equalized odds and statistical parity. Consistent with the experiments in Alghamdi *et al.* (2022), we

employ a gradient boosting classifier for all experiments. All results in this section are averaged over 10 trials.

Table 5.1: COMPAS dataset, statistical parity

Method	Accuracy	Statistical Parity Gap
OPT ($\varepsilon = 0.2$)	0.6798	0.2200
OPT ($\varepsilon = 1$)	0.6802	0.2346
OPT ($\varepsilon = 4$)	0.6804	0.2482
Non-Fair	0.6807	0.2553
Fair Projection - KL	0.6784	0.2201
Fair Projection - CE	0.6789	0.2197
Fair Projection - KL	0.6794	0.2345
Fair Projection - CE	0.6792	0.2348
Fair Projection - KL	0.6803	0.2486
Fair Projection - CE	0.6802	0.2480

Table 5.2: COMPAS dataset, mean equalized odds

Method	Accuracy	Mean Equalized Odds
OPT ($\varepsilon = 0.2$)	0.6798	0.1670
OPT ($\varepsilon = 1$)	0.6802	0.1780
OPT ($\varepsilon = 4$)	0.6804	0.1943
Non-Fair	0.6807	0.2106
Fair Projection - KL	0.6785	0.1675
Fair Projection - CE	0.6781	0.1664
Fair Projection - KL	0.6790	0.1785
Fair Projection - CE	0.6792	0.1776
Fair Projection - KL	0.6802	0.1941
Fair Projection - CE	0.6798	0.1938

For the COMPAS dataset, Tables 5.1 and 5.2 show that for a fixed level of statistical parity gap or mean equalized odds, OPT performs slightly better than Fair Projection - KL and Fair Projection - CE in terms of accuracy. For instance, OPT

Table 5.3: Adult dataset, statistical parity

Method	Accuracy	Statistical Parity Gap
OPT ($\varepsilon = 0.2$)	0.8609	0.1094
OPT ($\varepsilon = 1$)	0.8611	0.1099
OPT ($\varepsilon = 4$)	0.8611	0.1101
Non-Fair	0.8611	0.1151
Fair Projection - KL	0.8607	0.1090
Fair Projection - CE	0.8606	0.1092
Fair Projection - KL	0.8610	0.1101
Fair Projection - CE	0.8610	0.1101

Table 5.4: Adult dataset, mean equalized odds

Method	Accuracy	Mean Equalized Odds
OPT ($\varepsilon = 0.2$)	0.8609	0.0573
OPT ($\varepsilon = 1$)	0.8611	0.0580
OPT ($\varepsilon = 4$)	0.8611	0.0583
Non-Fair	0.8611	0.0668
Fair Projection - KL	0.8607	0.0573
Fair Projection - CE	0.8607	0.0573
Fair Projection - KL	0.8610	0.0584
Fair Projection - CE	0.8610	0.0582

with $\varepsilon = 1$ achieves an accuracy of 0.6802, while Fair Projection - KL and Fair Projection - CE attain accuracies of 0.6794 and 0.6792, respectively, with nearly the same statistical parity gap (Table 5.1). Similarly, they achieve accuracies of 0.6790 and 0.6792 with almost equal mean equalized odds (Table 5.2). Compared to the non-fair case, OPT improves fairness performance by more than 3% in statistical parity gap and more than 4% in mean equalized odds, with only a slight drop in accuracy. Note that Fair Projection applies specific post-processing for each fairness metric, which is why we provide two tables for each dataset corresponding to the results for mean equalized odds and statistical parity.

Similar trends can be observed in the Adult dataset, with a smaller difference between Fair Projection and OPT in terms of accuracy. Specifically, from Tables 5.3 and 5.4, we see that OPT can achieve a point on the trade-off curve of Fair Projection - CE and Fair Projection - KL with nearly identical accuracy and fairness, or with a slight accuracy gain at a fixed fairness level. Note that OPT requires significantly less runtime compared to Fair Projection, as the pre-processing step performed by the optimal LDP-based method is considerably faster than the post-processing optimization carried out by Fair Projection.

Chapter 6

Conclusion

6.1 Summary

In the first part of this thesis, we explored the intersection of central DP and fairness in binary classification, contributing to the development of a privacy-preserving fair classification algorithm. We first established a lower bound for the sum of prediction changes across a pair of subgroups under statistical parity and proposed Algorithm 1, which attains this bound without privacy constraints. Building on this foundation, we introduced Algorithm 2, a differentially private version that ensures both fairness and utility guarantees for the resulting classifier. Extensive experiments on the Adult and Credit Card datasets showed that our algorithm performs competitively in terms of accuracy compared to existing DP-Fair classification methods, particularly the DP-FERMI approach. For a given accuracy and privacy level, our method provides a stronger fairness guarantee, demonstrating the effectiveness of our approach in practical applications.

In the second part, we focused on the intersection of LDP and fairness, addressing

the challenge of minimizing data unfairness with LDP mechanisms. We formulated the problem of finding optimal LDP-based mechanisms for both binary and non-binary sensitive attributes, deriving a closed-form solution for the binary case and reformulating the non-binary case as a min-max linear fractional program which can be solved numerically. We demonstrated theoretically that reducing data unfairness with our LDP mechanisms leads to lower classification unfairness for certain types of classifiers, validating our objective to minimize data unfairness. Our empirical results across multiple datasets and fairness metrics further showed that our approach achieves similar utility levels as well-known LDP mechanisms while ensuring reduced unfairness post-classification. Additionally, our comparisons with the Fair Projection framework demonstrated that our optimal mechanism either matches or slightly surpasses Fair Projection's accuracy for fixed fairness levels, underscoring the potential of LDP as a simple yet powerful tool for mitigating data unfairness and enhancing classification fairness in machine learning.

6.2 Limitations and Future Directions

6.2.1 Central DP and Fairness

In Chapter 3, we introduced a post-processing algorithm for learning a binary classifier, $h^*_{\varepsilon',\delta',\text{Fair}}$, which guarantees (ε', δ')-DP while providing theoretical guarantees for both its utility and statistical parity gap. For fairness, we derived the following upper bound:

$$\Delta_{SP}(h^*_{\varepsilon',\delta',\text{Fair}}) \leq_\eta \frac{\log(4/\eta)}{n\theta\varepsilon_0} + \frac{\log(4/\eta)}{n\bar{\theta}\varepsilon_1} + \sqrt{\log(\frac{8}{\eta})\frac{1}{2n\theta}} + \sqrt{\log(\frac{8}{\eta})\frac{1}{2n\bar{\theta}}},$$

This upper bound on the unfairness of the learned classifier holds with high probability, ensuring that the classifier produced by Algorithm 2 will have bounded unfairness and thus meet a fairness criterion with high probability.

Empirical studies have shown that central DP can negatively impact the fairness of a learned classifier (Bagdasaryan *et al.*, 2019; Pujol *et al.*, 2020; Farrand *et al.*, 2020) (see Section 2.2). This raises a fundamental question: given privacy parameters (ε, δ), what is the minimum achievable unfairness for a classifier that satisfies (ε, δ)-DP while maintaining non-trivial utility? In other words, is it possible to derive a (probabilistic) lower bound on unfairness, expressed in terms of ε and δ, across all classifiers that meet (ε, δ)-DP and provide a meaningful utility? Answering this question would reveal the true accuracy-fairness-privacy trade-off and clarify how close current approaches are to achieving this balance, representing a fundamental step forward in understanding this critical relationship.

6.2.2 Local DP and Fairness

In Chapter 5, we formulated the problem of finding the optimal LDP-based pre-processing mechanism to minimize data unfairness. For a binary sensitive attribute, we derived the optimal mechanism as follows:

Given that $p_{1|0} \leq p_{1|1}$, the optimal LDP pre-processing mechanism (p^*, q^*) is:

$$\text{If} \quad p_0 < p_1, \quad (p^*, q^*) = \left(1 - \frac{e^{-\varepsilon}}{2}, \frac{1}{2}\right),$$

$$\text{If} \quad p_1 < p_0, \quad (p^*, q^*) = \left(\frac{1}{2}, 1 - \frac{e^{-\varepsilon}}{2}\right),$$

where p and q are the parameters of the general LDP mechanism as defined in (5.4.1)

(see Theorem 14). For the binary case, given the LDP parameter ε and the data distribution $(p_{1|0}, p_{1|1}, p_0, p_1)$, we can directly identify the optimal perturbation mechanism. However, for a non-binary sensitive attribute A, the problem of finding the optimal pre-processing mechanism becomes:

$$\min_{\mathbf{Q}} \max_{a\in[k]} \left| \frac{\sum_{\substack{j=1\\ j\neq a}}^{k} p_{1|j} p_j q_{ja} + p_{1|a} p_a q_{aa} - \sum_{\substack{j=1\\ j\neq a}}^{k} \Pr(Y=1) p_j q_{ja} - \Pr(Y=1) p_a q_{aa}}{\sum_{\substack{j=1\\ j\neq a}}^{k} \Pr(Y=1) p_j q_{ja} + \Pr(Y=1) p_a q_{aa}} \right|$$

$$\text{s.t.} \quad \Big(1 - \sum_{\substack{a=1\\ a\neq j}}^{k} q_{ja}\Big) - e^{\varepsilon} q_{ij} \leq 0 \qquad \forall i, j \in [k], \quad i \neq j$$

$$\sum_{\substack{a=1\\ a\neq i}}^{k} q_{ia} \leq 1 \qquad \forall i \in [k]$$

$$q_{ij} \geq 0, \quad 1 - \sum_{\substack{a=1\\ a\neq i}}^{k} q_{ia} \geq q_{ij}, \quad 1 - \sum_{\substack{a=1\\ a\neq j}}^{k} q_{ja} \geq q_{ij} \qquad \forall i, j \in [k], \quad i \neq j$$

$$\sum_{i=1}^{k} \Big(1 - \sum_{\substack{a=1\\ a\neq i}}^{k} q_{ia}\Big) p_i \geq 1 - \zeta$$

Here, the optimal mechanism can be obtained by solving a min-max linear fractional program. Although this approach effectively reduces unfairness, it may require substantial computational time when the sensitive attribute has a large support size. An interesting question for future research is whether, in the non-binary sensitive attribute case, we can reformulate the problem to obtain a closed-form solution, similar to the binary case, or reduce it to a convex optimization problem instead of a non-convex min-max linear fractional program.

Appendix A

Proofs of Chapter 3

Proof of Proposition 6. We have classifiers $h_0^* : \mathcal{X} \to \{0,1\}$ and $h_1^* : \mathcal{X} \to \{0,1\}$ trained on subgroups specified by the sensitive attribute A with values 0 and 1, respectively. We can combine classifiers h_0^* and h_1^* to have a group aware classifier $h^* : \mathcal{X} \times \{0,1\} \to \{0,1\}$. Basically, $h^*(x,0) = h_0^*(x)\ \forall x \in \mathcal{X}$ and $h^*(x,1) = h_1^*(x)\ \forall x \in \mathcal{X}$. Let $Y^* = h^*(X,A)$ and $\hat{Y} = \hat{h}(X,A)$. Then for $a \in \{0,1\}$, we have:

$$\begin{aligned} d_{TV}(\mu_a(Y^*), \mu_a(\hat{Y})) &= \left|\mu_a(Y^* = 1) - \mu_a(\hat{Y} = 1)\right| \\ &= \left|\mathbb{E}_{\mu_a^X}[h_a^*(X)] - \mathbb{E}_{\mu_a^X}[\hat{h}_a(X)]\right| \\ &\leq \mathbb{E}_{\mu_a^X}\left[\left|h_a^*(X) - \hat{h}_a(X)\right|\right] \\ &= \mathbb{P}_{\mu_a^X}\left(Y^* \neq \hat{Y}\right). \end{aligned} \tag{A.0.1}$$

Therefore, from (A.0.1) it follows:

$$d_{TV}(\mu_a(Y^*), \mu_a(\hat{Y})) \leq \mathbb{P}_{\mu_a^X}\left(Y^* \neq \hat{Y}\right).$$

We assumed $\hat{Y} = \hat{h}(X, A)$ satisfies γ statistical parity ($\Delta_{SP}(\hat{h}) \leq \gamma$). Thus, we have:

$$d_{TV}(\mu_0(\hat{Y}), \mu_1(\hat{Y})) = \Big|\mu_0(\hat{Y} = 1) - \mu_1(\hat{Y} = 1)\Big| = \Delta_{SP}(\hat{h}) \leq \gamma.$$

Since $d_{TV}(\cdot, \cdot)$ is symmetric and satisfies the triangle inequality, we have:

$$\begin{aligned} d_{TV}(\mu_0(Y^*), \mu_1(Y^*)) &\leq d_{TV}(\mu_0(Y^*), \mu_0(\hat{Y})) + d_{TV}(\mu_0(\hat{Y}), \mu_1(\hat{Y})) + d_{TV}(\mu_1(\hat{Y}), \mu_1(Y^*)) \\ &\leq d_{TV}(\mu_0(Y^*), \mu_0(\hat{Y})) + \gamma + d_{TV}(\mu_1(Y^*), \mu_1(\hat{Y})). \end{aligned} \tag{A.0.2}$$

Combining (A.0.2) with (A.0.1), we have:

$$\begin{aligned} d_{TV}(\mu_0(Y^*), \mu_1(Y^*)) &\leq \mathbb{P}_{\mu_0^X}\left(Y^* \neq \hat{Y}\right) + \mathbb{P}_{\mu_1^X}\left(Y^* \neq \hat{Y}\right) + \gamma \\ &= \mathbb{P}_{\mu_0^X}(\hat{h}_0(X) \neq h_0^*(X)) + \mathbb{P}_{\mu_1^X}(\hat{h}_1(X) \neq h_1^*(X)) + \gamma. \end{aligned}$$

Therefore, we can obtain:

$$\begin{aligned} &\Big|\mathbb{P}_{\mu_0^X}(h_0^*(X) = 1) - \mathbb{P}_{\mu_1^X}(h_1^*(X) = 1)\Big| \\ &\qquad \leq \mathbb{P}_{\mu_0^X}(\hat{h}_0(X) \neq h_0^*(X)) + \mathbb{P}_{\mu_1^X}(\hat{h}_1(X) \neq h_1^*(X)) + \gamma. \end{aligned}$$

Thus, we have:

$$\begin{aligned} &\mathbb{P}_{\mu_0^X}(\hat{h}_0(X) \neq h_0^*(X)) + \mathbb{P}_{\mu_1^X}(\hat{h}_1(X) \neq h_1^*(X)) \\ &\qquad \geq \Big|\mathbb{P}_{\mu_0^X}(h_0^*(X) = 1) - \mathbb{P}_{\mu_1^X}(h_1^*(X) = 1)\Big| - \gamma. \end{aligned}$$

which concludes the proof of Proposition 6. □

Proof of Theorem 7. Let s in Algorithm 1 be the realization of the random variable S.

$$\mathbb{P}_{\mu}\left(h^*_{\text{Fair}}(X,A)=1|A=0\right)=\mathbb{P}_{\mu_0^X}(h_0^*(X)=1)\mathbb{P}(S\leq\frac{\alpha+\beta}{2\alpha})=\alpha\left(\frac{\alpha+\beta}{2\alpha}\right)=\frac{\alpha+\beta}{2}.$$

$$\begin{aligned}\mathbb{P}_{\mu}\left(h^*_{\text{Fair}}(X,A)=1|A=1\right)&=\mathbb{P}_{\mu_1^X}(h_1^*(X)=1)+\mathbb{P}_{\mu_1^X}(h_1^*(X)=0)\mathbb{P}\left(S\leq\frac{\alpha-\beta}{2(1-\beta)}\right)\\&=\beta+(1-\beta)\left(\frac{\alpha-\beta}{2(1-\beta)}\right)\\&=\frac{\alpha+\beta}{2}.\end{aligned}$$

We have:

$$\Delta_{SP}(h^*_{\text{Fair}})=\left|\frac{\alpha+\beta}{2}-\frac{\alpha+\beta}{2}\right|=0.$$

Therefore, perfect statistical parity is satisfied. Now we show that h^*_{Fair} is optimal.

$$\mathbb{P}_{\mu_0^X}(h^*_{\text{Fair}0}(X)\neq h_0^*(X))=\mathbb{P}_{\mu_0^X}(h_0^*(X)=1)\mathbb{P}\left(S>\frac{\alpha+\beta}{2\alpha}\right)=\alpha\left(\frac{\alpha-\beta}{2\alpha}\right)=\frac{\alpha-\beta}{2}.$$

$$\begin{aligned}\mathbb{P}_{\mu_1^X}(h^*_{\text{Fair}1}(X)\neq h_1^*(X))&=\mathbb{P}_{\mu_1^X}(h_1^*(X)=0)\mathbb{P}\left(S\leq\frac{\alpha-\beta}{2(1-\beta)}\right)\\&=(1-\beta)\left(\frac{\alpha-\beta}{2(1-\beta)}\right)=\frac{\alpha-\beta}{2}.\end{aligned}$$

Thus, we have:

$$\begin{aligned}&\mathbb{P}_{\mu_0^X}(h^*_{\text{Fair}0}(X)\neq h_0^*(X))+\mathbb{P}_{\mu_1^X}(h^*_{\text{Fair}1}(X)\neq h_1^*(X))\\&\qquad=\frac{\alpha-\beta}{2}+\frac{\alpha-\beta}{2}=\alpha-\beta=|\alpha-\beta|\\&\qquad=\left|\mathbb{P}_{\mu_0^X}(h_0^*(X)=1)-\mathbb{P}_{\mu_1^X}(h_1^*(X)=1)\right|.\end{aligned}$$

Therefore, h^*_{Fair} satisfies perfect statistical parity and attains the utility lower bound of Proposition 6. □

Proof of Theorem 8.

- Every time we access the dataset to compute a query, it is essential to account for the privacy budget being consumed. Algorithm 2 accesses the training dataset to train the classifier $h^*_{\varepsilon,\delta} : \mathcal{X} \times \{0,1\} \to \{0,1\}$. Additionally, computing $\tilde{\alpha}$ and $\tilde{\beta}$ involves using the post-processing dataset. Learning $h^*_{\varepsilon,\delta} : \mathcal{X} \times \{0,1\} \to \{0,1\}$, from which we derive two classifiers $h^*_{\varepsilon,\delta,0}$ and $h^*_{\varepsilon,\delta,1}$, was conducted with privacy parameters (ε, δ). Computing $\tilde{\alpha}$ and $\tilde{\beta}$ utilizes the Laplace mechanism followed by post-processing (projection). Hence, these two mechanisms will satisfy ε_0-DP and ε_1-DP Dwork *et al.* (2014). By basic composition, we can conclude that $h^*_{\varepsilon',\delta',\text{Fair}}$ satisfies (ε', δ')-DP with $\varepsilon' = \varepsilon + \varepsilon_0 + \varepsilon_1$ and $\delta' = \delta$. Note that we assume the number of data points belonging to each subgroup in a dataset is public knowledge. Specifically, we assume that θn and $\bar{\theta} n$ are publicly known, and thus computing them does not consume any privacy budget.

- Let $\alpha = \mathbb{P}_{\mu_0^X}(h^*_{\varepsilon,\delta,0}(X) = 1)$ and $\beta = \mathbb{P}_{\mu_1^X}(h^*_{\varepsilon,\delta,1}(X) = 1)$. Let $\bar{\alpha} = \alpha + e_0$ and $\bar{\beta} = \beta + e_1$. To prove the claims in the theorem, we assume $\tilde{\alpha} \geq \tilde{\beta}$. For the case that $\tilde{\alpha} < \tilde{\beta}$, we will have the same results by symmetry. Also, let $L_0 \sim \text{Lap}\big(\frac{1}{n\theta\varepsilon_0}\big)$ and $L_1 \sim \text{Lap}\big(\frac{1}{n\bar{\theta}\varepsilon_1}\big)$. In Algorithm 2, we sample l_0 from L_0 and l_1 from L_1. Similar to the proof of Theorem 7, let s in Algorithm 2 be the realization of the random variable S. By definition, we have:

$$\Delta_{SP}(h^*_{\varepsilon',\delta',\text{Fair}}) = \Big|\mathbb{P}_{\mu_0^X}\left(h^*_{\varepsilon',\delta',\text{Fair}}(X,0) = 1\right) - \mathbb{P}_{\mu_1^X}\left(h^*_{\varepsilon',\delta',\text{Fair}}(X,1) = 1\right)\Big|.$$

We first compute $\mathbb{P}_{\mu_0^X}\left(h^*_{\varepsilon',\delta',\text{Fair}}(X,0)=1\right)$:

$$\begin{aligned}\mathbb{P}_{\mu_0^X}\left(h^*_{\varepsilon',\delta',\text{Fair}}(X,0)=1\right) &= \mathbb{P}_{\mu_0^X}(h^*_{\varepsilon,\delta,0}(X)=1)\mathbb{P}\left(S\leq\frac{\tilde{\alpha}+\tilde{\beta}}{2\tilde{\alpha}}\right)\\ &= \alpha\frac{\tilde{\alpha}+\tilde{\beta}}{2\tilde{\alpha}}\\ &= \frac{\alpha}{\tilde{\alpha}}\left(\frac{\tilde{\alpha}+\tilde{\beta}}{2}\right).\end{aligned}$$

We then compute $\mathbb{P}_{\mu_1^X}\left(h^*_{\varepsilon',\delta',\text{Fair}}(X,1)=1\right)$:

$$\begin{aligned}&\mathbb{P}_{\mu_1^X}\left(h^*_{\varepsilon',\delta',\text{Fair}}(X,1)=1\right)\\ &\quad= \mathbb{P}_{\mu_1^X}(h^*_{\varepsilon,\delta,1}(X)=1)+\mathbb{P}_{\mu_1^X}(h^*_{\varepsilon,\delta,1}(X)=0)\mathbb{P}\left(S\leq\frac{\tilde{\alpha}-\tilde{\beta}}{2(1-\tilde{\beta})}\right)\\ &\quad= \beta+(1-\beta)\left(\frac{\tilde{\alpha}-\tilde{\beta}}{2(1-\tilde{\beta})}\right)=\beta+\frac{(1-\beta)}{(1-\tilde{\beta})}\left(\frac{\tilde{\alpha}-\tilde{\beta}}{2}\right).\end{aligned}$$

For each realization of $\tilde{\alpha}$ and $\tilde{\beta}$, let $\bar{\alpha}=\alpha+e_0$, $\tilde{\alpha}=\bar{\alpha}+d_0$, $\bar{\beta}=\beta+e_1$, $\tilde{\beta}=\bar{\beta}+d_1$. Thus, we have:

$$\begin{aligned}\Delta_{SP}(h^*_{\varepsilon',\delta',\text{Fair}}) &= \left|\frac{\alpha}{\tilde{\alpha}}\left(\frac{\tilde{\alpha}+\tilde{\beta}}{2}\right)-\left[\beta+\frac{(1-\beta)}{(1-\tilde{\beta})}\left(\frac{\tilde{\alpha}-\tilde{\beta}}{2}\right)\right]\right|\\ &= \left|\frac{\tilde{\alpha}-e_0-d_0}{\tilde{\alpha}}\left(\frac{\tilde{\alpha}+\tilde{\beta}}{2}\right)-\left[\tilde{\beta}-e_1-d_1+\frac{(1-\tilde{\beta}+e_1+d_1)}{(1-\tilde{\beta})}\left(\frac{\tilde{\alpha}-\tilde{\beta}}{2}\right)\right]\right|\\ &= \left|\left(\frac{\tilde{\alpha}+\tilde{\beta}}{2}\right)-(e_0+d_0)\left(\frac{\tilde{\alpha}+\tilde{\beta}}{2\tilde{\alpha}}\right)-\tilde{\beta}+e_1+d_1\right.\\ &\quad\left.-\left(\frac{\tilde{\alpha}-\tilde{\beta}}{2}\right)-(e_1+d_1)\left(\frac{\tilde{\alpha}-\tilde{\beta}}{2(1-\tilde{\beta})}\right)\right|\\ &= \left|(e_1+d_1)\left(1-\frac{\tilde{\alpha}-\tilde{\beta}}{2(1-\tilde{\beta})}\right)-(e_0+d_0)\left(\frac{\tilde{\alpha}+\tilde{\beta}}{2\tilde{\alpha}}\right)\right|\end{aligned}$$

$$\leq \left|(e_1 + d_1)\left(1 - \frac{\tilde{\alpha} - \tilde{\beta}}{2(1-\tilde{\beta})}\right)\right| + \left|(e_0 + d_0)\left(\frac{\tilde{\alpha} + \tilde{\beta}}{2\tilde{\alpha}}\right)\right|$$
$$\leq |(e_1 + d_1)| + |(e_0 + d_0)|$$
$$\leq |e_0| + |e_1| + |d_0| + |d_1|. \tag{A.0.3}$$

The last line follows from the fact that $0 \leq \tilde{\alpha} \leq 1$, $0 \leq \tilde{\beta} \leq 1$, and $\tilde{\alpha} \geq \tilde{\beta}$.

We know that $|e_0| = |\bar{\alpha} - \alpha| = \left|\frac{1}{n\theta} \sum_{\substack{i=1 \\ A_i=0}}^{n} h^*_{\varepsilon,\delta,0}(X_i) - \mathbb{P}_{\mu_0^X}(h^*_{\varepsilon,\delta,0}(X) = 1)\right|$. From Hoeffding's inequality, we know that if $X_1, X_2, \ldots, X_n$ are i.i.d. random variables in $[0, 1]$, then:

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right| \geq t\right] \leq 2e^{-2nt^2}.$$

Therefore, we can conclude that:

$$\mathbb{P}\left[\left|\frac{1}{n\theta}\sum_{\substack{i=1 \\ A_i=0}}^{n} h^*_{\varepsilon,\delta,0}(X_i) - \mathbb{P}_{\mu_0^X}(h^*_{\varepsilon,\delta,0}(X) = 1)\right| \geq t\right] \leq 2e^{-2n\theta t^2},$$

Which means:

$$\mathbb{P}\left[\left|\frac{1}{n\theta}\sum_{\substack{i=1 \\ A_i=0}}^{n} h^*_{\varepsilon,\delta,0}(X_i) - \mathbb{P}_{\mu_0^X}(h^*_{\varepsilon,\delta,0}(X) = 1)\right| \geq \sqrt{\frac{1}{2n\theta}\log(\frac{2}{\eta})}\right] \leq \eta.$$

Thus, we have:

$$|e_0| \leq_\eta \sqrt{\frac{1}{2n\theta}\log(\frac{2}{\eta})}. \tag{A.0.4}$$

Similarly, it follows that:

$$|e_1| \leq_\eta \sqrt{\frac{1}{2n\bar{\theta}} \log(\frac{2}{\eta})}. \tag{A.0.5}$$

Let l_0 and l_1 be the realization of the Laplace noises in Algorithm 2. We know that $|d_0| \leq |l_0|$ and $|d_1| \leq |l_1|$ since d_0 and d_1 are computed after projection. On the other hand, from Dwork *et al.* (2014), we know that if $L \sim \text{Lap}(\frac{\Delta_1^q}{\varepsilon})$, then:

$$\mathbb{P}\left[|L| \geq \left(\log \frac{1}{\eta}\right)\left(\frac{\Delta_1^q}{\varepsilon}\right)\right] \leq \eta.$$

where Δ_1^q is the ℓ_1-sensitivity of the query to which we add noise.

Given $L_0 \sim \text{Lap}\left(\frac{1}{n\theta\varepsilon_0}\right)$ and $L_1 \sim \text{Lap}\left(\frac{1}{n\bar{\theta}\varepsilon_1}\right)$, we have:
$|L_0| \leq_\eta \left(\log \frac{1}{\eta}\right)\left(\frac{1}{n\theta\varepsilon_0}\right)$ and $|L_1| \leq_\eta \left(\log \frac{1}{\eta}\right)\left(\frac{1}{n\bar{\theta}\varepsilon_1}\right)$. Comparing these two inequalities with (A.0.4) and (A.0.5), we can conclude that with probability at least $(1-\eta)^4$, we have:

$$\begin{aligned}
&|L_0| + |L_1| + |e_0| + |e_1| \\
&\quad \leq \left[\left(\log \frac{1}{\eta}\right)\left(\frac{1}{n\theta\varepsilon_0}\right) + \left(\log \frac{1}{\eta}\right)\left(\frac{1}{n\bar{\theta}\varepsilon_1}\right) + \sqrt{\frac{1}{2n\theta}\log(\frac{2}{\eta})} + \sqrt{\frac{1}{2n\bar{\theta}}\log(\frac{2}{\eta})}\right].
\end{aligned}$$

Since $(1-\eta)^4 \geq 1 - 4\eta$ for $0 \leq \eta \leq 1$, we have:

$$\begin{aligned}
&|L_0| + |L_1| + |e_0| + |e_1| \\
&\quad \leq_{4\eta} \left[\left(\log \frac{1}{\eta}\right)\left(\frac{1}{n\theta\varepsilon_0}\right) + \left(\log \frac{1}{\eta}\right)\left(\frac{1}{n\bar{\theta}\varepsilon_1}\right) + \sqrt{\frac{1}{2n\theta}\log(\frac{2}{\eta})} + \sqrt{\frac{1}{2n\bar{\theta}}\log(\frac{2}{\eta})}\right].
\end{aligned}$$

Equivalently, we have:

$$|L_0| + |L_1| + |e_0| + |e_1| \leq_\eta \left[\left(\log\frac{4}{\eta}\right)\left(\frac{1}{n\theta\varepsilon_0}\right) + \left(\log\frac{4}{\eta}\right)\left(\frac{1}{n\bar{\theta}\varepsilon_1}\right) + \sqrt{\frac{1}{2n\theta}\log(\frac{8}{\eta})} + \sqrt{\frac{1}{2n\bar{\theta}}\log(\frac{8}{\eta})}\right]. \tag{A.0.6}$$

Combining (A.0.3) and (A.0.6), it can be shown:

$$\Delta_{SP}(h^*_{\varepsilon',\delta',\text{Fair}}) \leq_\eta \left[\left(\log\frac{4}{\eta}\right)\left(\frac{1}{n\theta\varepsilon_0}\right) + \left(\log\frac{4}{\eta}\right)\left(\frac{1}{n\bar{\theta}\varepsilon_1}\right) + \sqrt{\frac{1}{2n\theta}\log(\frac{8}{\eta})} + \sqrt{\frac{1}{2n\bar{\theta}}\log(\frac{8}{\eta})}\right].$$

- We have:

$$\begin{aligned}
&\mathbb{P}_{\mu_0^X}(h^*_{\varepsilon',\delta',\text{Fair}_0}(X) \neq h^*_{\varepsilon,\delta,0}(X)) + \mathbb{P}_{\mu_1^X}(h^*_{\varepsilon',\delta',\text{Fair}_1}(X) \neq h^*_{\varepsilon,\delta,1}(X)) \\
&\quad= \mathbb{P}_{\mu_0^X}\Big((h^*_{\varepsilon,\delta,0}(X) = 1) \text{ and } (h^*_{\varepsilon',\delta',\text{Fair}_0}(X) = 0)\Big) \\
&\quad+ \mathbb{P}_{\mu_1^X}\Big((h^*_{\varepsilon,\delta,1}(X) = 0) \text{ and } (h^*_{\varepsilon',\delta',\text{Fair}_1}(X) = 1)\Big) \\
&\quad= \alpha\left(\frac{\tilde{\alpha}-\tilde{\beta}}{2\tilde{\alpha}}\right) + (1-\beta)\left(\frac{\tilde{\alpha}-\tilde{\beta}}{2(1-\tilde{\beta})}\right) = \frac{\alpha}{\tilde{\alpha}}\left(\frac{\tilde{\alpha}-\tilde{\beta}}{2}\right) + \frac{(1-\beta)}{1-\tilde{\beta}}\left(\frac{\tilde{\alpha}-\tilde{\beta}}{2}\right) \\
&\quad= \frac{\tilde{\alpha}-e_0-d_0}{\tilde{\alpha}}\left(\frac{\tilde{\alpha}-\tilde{\beta}}{2}\right) + \frac{(1-\tilde{\beta}+e_1+d_1)}{1-\tilde{\beta}}\left(\frac{\tilde{\alpha}-\tilde{\beta}}{2}\right) \\
&\quad= (\tilde{\alpha}-\tilde{\beta}) + (e_1+d_1)\left(\frac{\tilde{\alpha}-\tilde{\beta}}{2(1-\tilde{\beta})}\right) - (e_0+d_0)\left(\frac{\tilde{\alpha}-\tilde{\beta}}{2\tilde{\alpha}}\right).
\end{aligned}$$

Therefore, we have:

$$\begin{aligned}&\mathbb{P}_{\mu_0^X}(h^*_{\varepsilon',\delta',\text{Fair}_0}(X) \neq h^*_{\varepsilon,\delta,0}(X)) + \mathbb{P}_{\mu_1^X}(h^*_{\varepsilon',\delta',\text{Fair}_1}(X) \neq h^*_{\varepsilon,\delta,1}(X)) - \text{err}^*(h^*_{\varepsilon,\delta,0}, h^*_{\varepsilon,\delta,1})\\ &\quad= (\tilde{\alpha} - \tilde{\beta}) + (e_1 + d_1)\left(\frac{\tilde{\alpha} - \tilde{\beta}}{2(1-\tilde{\beta})}\right) - (e_0 + d_0)\left(\frac{\tilde{\alpha} - \tilde{\beta}}{2\tilde{\alpha}}\right) - \text{err}^*(h^*_{\varepsilon,\delta,0}, h^*_{\varepsilon,\delta,1}).\end{aligned}$$

From previous part, we know that:

$$\Delta_{SP}(h^*_{\varepsilon',\delta',\text{Fair}}) \leq |e_0| + |e_1| + |d_0| + |d_1|.$$

From Proposition 6, we know that if $\Delta_{SP}(\hat{h}) \leq \gamma$, then:

$$\begin{aligned}&\mathbb{P}_{\mu_0^X}(\hat{h}_0(X) \neq h^*_{\varepsilon,\delta,0}(X)) + \mathbb{P}_{\mu_1^X}(\hat{h}_1(X) \neq h^*_{\varepsilon,\delta,1}(X))\\ &\qquad\qquad\geq \left|\mathbb{P}_{\mu_0^X}(h^*_{\varepsilon,\delta,0}(X) = 1) - \mathbb{P}_{\mu_1^X}(h^*_{\varepsilon,\delta,1}(X) = 1)\right| - \gamma.\end{aligned}$$

For all classifiers $\hat{h} : \mathcal{X} \times \{0,1\} \to \{0,1\}$ that satisfy $\Delta_{SP}(\hat{h}) \leq \Delta_{SP}(h^*_{\varepsilon',\delta',\text{Fair}})$, we have $\Delta_{SP}(\hat{h}) \leq |e_0| + |e_1| + |d_0| + |d_1|$. Thus, for all those classifiers we have:

$$\begin{aligned}&\mathbb{P}_{\mu_0^X}(\hat{h}_0(X) \neq h^*_{\varepsilon,\delta,0}(X)) + \mathbb{P}_{\mu_1^X}(\hat{h}_1(X) \neq h^*_{\varepsilon,\delta,1})\\ &\quad\geq \left|\mathbb{P}_{\mu_0^X}(h^*_{\varepsilon,\delta,0}(X) = 1) - \mathbb{P}_{\mu_1^X}(h^*_{\varepsilon,\delta,1}(X) = 1)\right| - \big(|e_0| + |e_1| + |d_0| + |d_1|\big)\\ &\quad= |\alpha - \beta| - \big(|e_0| + |e_1| + |d_0| + |d_1|\big).\end{aligned}$$

Therefore, by definition of $\text{err}^*(h^*_{\varepsilon,\delta,0}, h^*_{\varepsilon,\delta,1})$, we have:

$$\text{err}^*(h^*_{\varepsilon,\delta,0}, h^*_{\varepsilon,\delta,1}) \geq |\alpha - \beta| - (|e_0| + |e_1| + |d_0| + |d_1|)\,.$$

Therefore, it follows:

$$
\begin{aligned}
&\mathbb{P}_{\mu_0^X}(h^*_{\varepsilon',\delta',\mathrm{Fair}_0}(X) \neq h^*_{\varepsilon,\delta,0}(X)) + \mathbb{P}_{\mu_1^X}(h^*_{\varepsilon',\delta',\mathrm{Fair}_1}(X) \neq h^*_{\varepsilon,\delta,1}(X)) - \mathrm{err}^*(h^*_{\varepsilon,\delta,0}, h^*_{\varepsilon,\delta,1}) \\
&\qquad \leq (\tilde{\alpha} - \tilde{\beta}) + (e_1 + d_1)\left(\frac{\tilde{\alpha} - \tilde{\beta}}{2(1-\tilde{\beta})}\right) - (e_0 + d_0)\left(\frac{\tilde{\alpha} - \tilde{\beta}}{2\tilde{\alpha}}\right) \\
&\qquad - |\alpha - \beta| + (|e_0| + |e_1| + |d_0| + |d_1|) \\
&\qquad \leq \frac{1}{2}|e_1 + d_1| + \frac{1}{2}|e_0 + d_0| + (\tilde{\alpha} - \tilde{\beta}) - |\alpha - \beta| + (|e_0| + |e_1| + |d_0| + |d_1|) \\
&\qquad \leq (\tilde{\alpha} - \tilde{\beta}) - |\alpha - \beta| + \frac{3}{2}(|e_0| + |e_1| + |d_0| + |d_1|) \\
&\qquad = (\tilde{\alpha} - \tilde{\beta}) - |\tilde{\alpha} - e_0 - d_0 - \tilde{\beta} + e_1 + d_1| + \frac{3}{2}(|e_0| + |e_1| + |d_0| + |d_1|) \\
&\qquad \leq \frac{5}{2}(|e_0| + |e_1| + |d_0| + |d_1|)\,.
\end{aligned}
$$

By the same argument of the second part, we can conclude that:

$$
\begin{aligned}
&\mathbb{P}_{\mu_0^X}(h^*_{\varepsilon',\delta',\mathrm{Fair}_0}(X) \neq h^*_{\varepsilon,\delta,0}(X)) + \mathbb{P}_{\mu_1^X}(h^*_{\varepsilon',\delta',\mathrm{Fair}_1}(X) \neq h^*_{\varepsilon,\delta,1}(X)) \\
&\quad \leq_\eta \mathrm{err}^*(h^*_{\varepsilon,\delta,0}, h^*_{\varepsilon,\delta,1}) \\
&\quad + \frac{5}{2}\left(\left(\log\frac{4}{\eta}\right)\left(\frac{1}{n\theta\varepsilon_0}\right) + \left(\log\frac{4}{\eta}\right)\left(\frac{1}{n\bar{\theta}\varepsilon_1}\right) + \sqrt{\frac{1}{2n\theta}\log(\frac{8}{\eta})} + \sqrt{\frac{1}{2n\bar{\theta}}\log(\frac{8}{\eta})}\right).
\end{aligned}
$$

□

Proof of Proposition 9. Let $\alpha = \mathbb{P}_{\mu_0^X}(h^*_{\varepsilon,\delta,0}(X) = 1)$ and $\beta = \mathbb{P}_{\mu_1^X}(h^*_{\varepsilon,\delta,1}(X) = 1)$. Also, let $\bar{\alpha} = \alpha + e_0$ and $\bar{\beta} = \beta + e_1$. Let $L_0 \sim \mathrm{Lap}(\frac{1}{n\theta\varepsilon_0})$ and $L_1 \sim \mathrm{Lap}(\frac{1}{n\bar{\theta}\varepsilon_1})$. We sample l_0 from L_0 and l_1 from L_1. In fact, l_0 and l_1 are realizations of the Laplace noise. From proof of Theorem 8, we know that for each realization of the noise we

have:

$$\Delta_{SP}(h^*_{\varepsilon',\delta',\text{Fair}}) \leq [|e_0| + |e_1| + |l_0| + |l_1|]\,.$$

And

$$\left[\mathbb{P}_{\mu_0^X}(h^*_{\varepsilon',\delta',\text{Fair}_0}(X) \neq h^*_{\varepsilon,\delta,0}(X)) + \mathbb{P}_{\mu_1^X}(h^*_{\varepsilon',\delta',\text{Fair}_1}(X) \neq h^*_{\varepsilon,\delta,1}(X)) - \text{err}^*(h^*_{\varepsilon,\delta,0}, h^*_{\varepsilon,\delta,1})\right] \leq \frac{5}{2}\left[(|e_0| + |e_1| + |l_0| + |l_1|)\right].$$

We have:

$$\mathbb{E}\left[\Delta_{SP}(h^*_{\varepsilon',\delta',\text{Fair}})\right] \leq \mathbb{E}\left[|e_0| + |e_1| + |L_0| + |L_1|\right]. \tag{A.0.7}$$

And

$$\mathbb{E}\left[\mathbb{P}_{\mu_0^X}(h^*_{\varepsilon',\delta',\text{Fair}_0}(X) \neq h^*_{\varepsilon,\delta,0}(X)) + \mathbb{P}_{\mu_1^X}(h^*_{\varepsilon',\delta',\text{Fair}_1}(X) \neq h^*_{\varepsilon,\delta,1}(X)) - \text{err}^*(h^*_{\varepsilon,\delta,0}, h^*_{\varepsilon,\delta,1})\right] \leq \frac{5}{2}\mathbb{E}\left[(|e_0| + |e_1| + |L_0| + |L_1|)\right]. \tag{A.0.8}$$

Where the expectations are over the randomness of the Laplace noise and randomness of the approximation of α and β with finite samples. For $L \sim \text{Lap}(\frac{\Delta_1^q}{\varepsilon})$, we have $\mathbb{E}\left[|L|\right] = \frac{\Delta_1^q}{\varepsilon}$. Since $L_0 \sim \text{Lap}\left(\frac{1}{n\theta\varepsilon_0}\right)$ and $L_1 \sim \text{Lap}\left(\frac{1}{n\bar{\theta}\varepsilon_1}\right)$, we have:

$$\mathbb{E}\left[|L_0| + |L_1|\right] = \left(\frac{1}{n\theta\varepsilon_0}\right) + \left(\frac{1}{n\bar{\theta}\varepsilon_1}\right). \tag{A.0.9}$$

Now, we want to find an upper bound for $\mathbb{E}[|e_0|]$ and $\mathbb{E}[|e_1|]$. We have:

$$\begin{aligned}
\mathbb{E}[|e_0|]^2 &= \mathbb{E}\left[\left|\frac{1}{n\theta}\sum_{\substack{i=1\\A_i=0}}^{n} h^*_{\varepsilon,\delta,0}(X_i) - \mathbb{P}_{\mu_0^X}(h^*_{\varepsilon,\delta,0}(X)=1)\right|\right]^2 \\
&\leq \mathbb{E}\left[\left|\frac{1}{n\theta}\sum_{\substack{i=1\\A_i=0}}^{n} h^*_{\varepsilon,\delta,0}(X_i) - \mathbb{P}_{\mu_0^X}(h^*_{\varepsilon,\delta,0}(X)=1)\right|^2\right] \\
&= \mathbb{E}\left[\left(\frac{1}{n\theta}\sum_{\substack{i=1\\A_i=0}}^{n} h^*_{\varepsilon,\delta,0}(X_i) - \mathbb{E}\left[\frac{1}{n\theta}\sum_{\substack{i=1\\A_i=0}}^{n} h^*_{\varepsilon,\delta,0}(X_i)\right]\right)^2\right] \\
&= \mathrm{Var}\left(\frac{1}{n\theta}\sum_{\substack{i=1\\A_i=0}}^{n} h^*_{\varepsilon,\delta,0}(X_i)\right) \\
&= \frac{1}{\theta^2 n^2}\mathrm{Var}\left(\sum_{\substack{i=1\\A_i=0}}^{n} h^*_{\varepsilon,\delta,0}(X_i)\right) \\
&= \frac{1}{n\theta}\mathrm{Var}\left(h^*_{\varepsilon,\delta,0}(X_i)\right) \\
&\leq \frac{1}{4n\theta}.
\end{aligned}$$

The inequality in the second line is due to the property that for any random variable Z, we have $\mathbb{E}[|Z|]^2 \leq \mathbb{E}[Z^2]$, and the inequality in the last line follows from $\mathrm{Var}(\text{Bernoulli}(p)) = p(1-p) \leq \frac{1}{4}$ for $0 \leq p \leq 1$.

Similarly, it follows that:

$$\mathbb{E}[|e_1|]^2 \leq \frac{1}{4n\bar{\theta}}.$$

Therefore, it can be shown that:

$$\mathbb{E}[|e_0|] + \mathbb{E}[|e_1|] \leq \sqrt{\frac{1}{4n\theta}} + \sqrt{\frac{1}{4n\bar{\theta}}}. \tag{A.0.10}$$

Combining (A.0.7), (A.0.9), and (A.0.10), we have:

$$\mathbb{E}\left[\Delta_{SP}(h^*_{\varepsilon',\delta',\text{Fair}})\right] \leq \frac{1}{n\theta\varepsilon_0} + \frac{1}{n\bar{\theta}\varepsilon_1} + \sqrt{\frac{1}{4n\theta}} + \sqrt{\frac{1}{4n\bar{\theta}}}.$$

With the same argument and Using (A.0.8), it can be shown that:

$$\mathbb{E}\left[\mathbb{P}_{\mu_0^X}(h^*_{\varepsilon',\delta',\text{Fair}_0}(X) \neq h^*_{\varepsilon,\delta,0}(X)) + \mathbb{P}_{\mu_1^X}(h^*_{\varepsilon',\delta',\text{Fair}_1}(X) \neq h^*_{\varepsilon,\delta,1}(X))\right] \leq$$
$$\mathbb{E}\left[\text{err}^*(h^*_{\varepsilon,\delta,0}, h^*_{\varepsilon,\delta,1})\right] + \frac{5}{2}\left(\frac{1}{n\theta\varepsilon_0} + \frac{1}{n\bar{\theta}\varepsilon_1} + \sqrt{\frac{1}{4n\theta}} + \sqrt{\frac{1}{4n\bar{\theta}}}\right).$$

□

Appendix B

Proofs of Chapter 4 and 5

Proof of Lemma 13. To simplify notation, we define the variables p_j, $p_{1|j}$ for $j \in [k]$ as follows:

$$p_j = \Pr(A = j) \quad \forall j \in [k], \qquad p_{1|j} = \Pr(Y = 1 \,|\, A = j) \quad \forall j \in [k].$$

We know:

$$\Delta'(D) := \max_{a,a' \in [k]} \left| \Pr(Y = 1 \mid A = a) - \Pr(Y = 1 \mid A = a') \right| = \max_{a \in [k]} p_{1|a} - \min_{a \in [k]} p_{1|a}.$$

Proof Sketch: W.L.O.G., we assume $\max_{a \in [k]} p_{1|a} = p_{1|k}$ and $\min_{a \in [k]} p_{1|a} = p_{1|1}$. Therefore, we have $\Delta'(D) = p_{1|k} - p_{1|1}$. Similarly, we can express $\Delta'(D^{\varepsilon}_{GRR})$ as:

$$\Delta'(D^{\varepsilon}_{GRR}) = \max_{a \in [k]} \Pr(Y = 1 \,|\, Z = a) - \min_{a \in [k]} \Pr(Y = 1 \,|\, Z = a).$$

For each $a \in [k]$, we prove $p_{1|1} \leq \Pr(Y = 1 \,|\, Z = a) \leq p_{1|k}$. This means that $\max_{a \in [k]} \Pr(Y = 1 \,|\, Z = a) \leq p_{1|k}$ and $p_{1|1} \leq \min_{a \in [k]} \Pr(Y = 1 \,|\, Z = a)$, concluding

$\Delta'(D^{\varepsilon}_{GRR}) \leq \Delta'(D)$.

$\forall a \in [k]$, We have:

$$\begin{aligned}
\Pr(Y=1 \mid Z=a) &= \frac{\Pr(Y=1, Z=a)}{\Pr(Z=a)} \\
&= \frac{\sum_{j=1}^{k} \Pr(Y=1, Z=a, A=j)}{\sum_{j=1}^{k} \Pr(Z=a, A=j)} \\
&= \frac{\sum_{j=1}^{k} \Pr(Y=1, A=j) \Pr(Z=a \mid A=j, Y=1)}{\sum_{j=1}^{k} \Pr(Z=a \mid A=j) \Pr(A=j)} \\
&= \frac{\sum_{j=1}^{k} \Pr(Y=1, A=j) \Pr(Z=a \mid A=j)}{\sum_{j=1}^{k} \Pr(Z=a \mid A=j) \Pr(A=j)} \\
&= \frac{\sum_{j=1}^{k} \Pr(Y=1 \mid A=j) \Pr(A=j) \Pr(Z=a \mid A=j)}{\sum_{j=1}^{k} \Pr(Z=a \mid A=j) \Pr(A=j)} \\
&= \frac{p_{1|a} p_a \pi + \sum_{\substack{j \in [k] \\ j \neq a}} p_{1|j} p_j \bar{\pi}}{p_a \pi + \sum_{\substack{j \in [k] \\ j \neq a}} p_j \bar{\pi}}.
\end{aligned}$$

We know $p_{1|1} \leq p_{1|j} \leq p_{1|k}$ for all values of $j \in [k]$. Therefore, it can be obtained that:

$$\frac{p_{1|1}p_a\pi + \sum\limits_{\substack{j\in[k]\\ j\neq a}} p_{1|1}p_j\bar{\pi}}{p_a\pi + \sum\limits_{\substack{j\in[k]\\ j\neq a}} p_j\bar{\pi}} \leq \frac{p_{1|a}p_a\pi + \sum\limits_{\substack{j\in[k]\\ j\neq a}} p_{1|j}p_j\bar{\pi}}{p_a\pi + \sum\limits_{\substack{j\in[k]\\ j\neq a}} p_j\bar{\pi}} \leq \frac{p_{1|k}p_a\pi + \sum\limits_{\substack{j\in[k]\\ j\neq a}} p_{1|k}p_j\bar{\pi}}{p_a\pi + \sum\limits_{\substack{j\in[k]\\ j\neq a}} p_j\bar{\pi}}.$$

Thus, we have:

$$p_{1|1} \leq \frac{p_{1|a}p_a\pi + \sum\limits_{\substack{j\in[k]\\ j\neq a}} p_{1|j}p_j\bar{\pi}}{p_a\pi + \sum\limits_{\substack{j\in[k]\\ j\neq a}} p_j\bar{\pi}} \leq p_{1|k} \quad \longrightarrow \quad p_{1|1} \leq \Pr(Y=1 \,|\, Z=a) \leq p_{1|k} \quad \forall a \in [k].$$

It follows that:

$$\begin{aligned}\Delta'(D^{\varepsilon}_{GRR}) &= \max_{a\in[k]} \Pr(Y=1 \,|\, Z=a) - \min_{a\in[k]} \Pr(Y=1 \,|\, Z=a)\\ &\leq p_{1|k} - p_{1|1}\\ &= \Delta'(D).\end{aligned}$$

Now the proof is complete. □

Proof of Theorem 14. To simplify notation, we define the variables p_j, $p_{1|j}$ for $j \in \{0,1\}$ as follows:

$$p_j = \Pr(A=j) \quad \forall j \in \{0,1\}, \qquad p_{1|j} = \Pr(Y=1 \,|\, A=j) \quad \forall j \in \{0,1\}.$$

We know:

$$\Delta'(D) = \max_{a,a' \in \{0,1\}} \Big| \Pr(Y = 1 \mid A = a) - \Pr(Y = 1 \mid A = a') \Big|.$$

Similarly, we can express $\Delta'(D^{\varepsilon}_{M})$ as:

$$\Delta'(D^{\varepsilon}_{M}) = \max_{a,a' \in \{0,1\}} \Big| \Pr(Y = 1 \mid Z = a) - \Pr(Y = 1 \mid Z = a') \Big|.$$

Proof Sketch: We assume that $\max_{a \in \{0,1\}} p_{1|a} = p_{1|1}$ and $\min_{a \in \{0,1\}} p_{1|a} = p_{1|0}$. Therefore, we have $\Delta'(D) = p_{1|1} - p_{1|0}$. Proof consists of three steps. In the first step, we find the set of feasible values of p and q such that the mechanism M satisfies ε-LDP. In the second step, we prove that $\Pr(Y = 1 \,|\, Z = 0) \leq \Pr(Y = 1 \,|\, Z = 1)$. Finally, in the third step, we determine the values of p and q that minimize $\frac{\Delta'(D^{\varepsilon}_{M})}{\Delta'(D)}$.

Step 1: We start by assuming that $\frac{1}{2} \leq p \leq 1$ and $\frac{1}{2} \leq q \leq 1$. This is important because simply replacing the sensitive attribute of each individual randomly with 0 or 1 (with a misreporting probability of $\frac{1}{2}$) would satisfy 0-LDP, but it would severely compromise utility. Therefore, to ensure non-trivial utility, it is necessary that $\frac{1}{2} \leq p \leq 1$ and $\frac{1}{2} \leq q \leq 1$. Also, by definition of an ε-LDP mechanism, we have:

$$(1 - p) - e^{\varepsilon} q \leq 0$$

$$p - e^{\varepsilon}(1 - q) \leq 0$$

$$q - e^{\varepsilon}(1 - p) \leq 0$$

$$(1 - q) - e^{\varepsilon} p \leq 0.$$

From $\frac{1}{2} \leq p \leq 1$ and $\frac{1}{2} \leq q \leq 1$ directly follows that $(1 - p) - e^{\varepsilon} q \leq 0$ and

$(1-q) - e^{\varepsilon}p \leq 0$. However, to satisfy $p - e^{\varepsilon}(1-q) \leq 0$ and $q - e^{\varepsilon}(1-p) \leq 0$ we should have:

$$p - e^{\varepsilon}(1-q) \leq 0 \implies p \leq e^{\varepsilon}(1-q).$$

And

$$q - e^{\varepsilon}(1-p) \leq 0 \implies q \leq e^{\varepsilon}(1-p).$$

In conclusion, any mechanism M that guarantees ε-LDP and non-trivial utility should satisfy the following inequalities:

$$\frac{1}{2} \leq p \leq 1, \quad \frac{1}{2} \leq q \leq 1, \quad p \leq e^{\varepsilon}(1-q), \quad q \leq e^{\varepsilon}(1-p).$$

Step 2: $\forall a \in \{0,1\}$, We have:

$$\begin{aligned}
\Pr(Y=1 \,|\, Z=a) &= \frac{\Pr(Y=1, Z=a)}{\Pr(Z=a)} \\
&= \frac{\sum_{j=0}^{1} \Pr(Y=1, Z=a, A=j)}{\sum_{j=0}^{1} \Pr(Z=a, A=j)} \\
&= \frac{\sum_{j=0}^{1} \Pr(Y=1, A=j) \Pr(Z=a \,|\, A=j, Y=1)}{\sum_{j=0}^{1} \Pr(Z=a \,|\, A=j) \Pr(A=j)} \\
&= \frac{\sum_{j=0}^{1} \Pr(Y=1, A=j) \Pr(Z=a \,|\, A=j)}{\sum_{j=0}^{1} \Pr(Z=a \,|\, A=j) \Pr(A=j)}
\end{aligned}$$

$$= \frac{\sum_{j=0}^{1} \Pr(Y = 1 \mid A = j) \Pr(A = j) \Pr(Z = a \mid A = j)}{\sum_{j=0}^{1} \Pr(Z = a \mid A = j) \Pr(A = j)}.$$

Therefore, we have:

$$\Pr(Y = 1 \mid Z = 0) = \frac{p_{1|0}p_0p + p_{1|1}p_1(1-q)}{p_0p + p_1(1-q)},$$

And

$$\Pr(Y = 1 \mid Z = 1) = \frac{p_{1|0}p_0(1-p) + p_{1|1}p_1(q)}{p_0(1-p) + p_1q}.$$

To demonstrate that $\Pr(Y = 1 \mid Z = 0) \leq \Pr(Y = 1 \mid Z = 1)$, we need to prove the following inequality:

$$\Big(p_0(1-p)+p_1q\Big)\Big(p_{1|0}p_0p+p_{1|1}p_1(1-q)\Big) \leq \Big(p_0p+p_1(1-q)\Big)\Big(p_{1|0}p_0(1-p)+p_{1|1}p_1(q)\Big).$$

Expanding both sides of the inequality, we have:

$$p_{1|0}(p_0)^2p(1-p) + p_{1|1}p_0p_1(1-p)(1-q) + p_{1|0}p_0p_1pq + p_{1|1}(p_1)^2q(1-q) \leq$$
$$p_{1|0}(p_0)^2p(1-p) + p_{1|1}p_0p_1pq + p_{1|0}p_0p_1(1-q)(1-p) + p_{1|1}(p_1)^2q(1-q).$$

Simplifying this, we get:

$$p_{1|1}p_0p_1(1-p)(1-q) + p_{1|0}p_0p_1pq \leq p_{1|1}p_0p_1pq + p_{1|0}p_0p_1(1-p)(1-q).$$

This simplifies further to:

$$p_{1|1}(1-p)(1-q) + p_{1|0}pq \leq p_{1|1}pq + p_{1|0}(1-p)(1-q).$$

Rearranging, we obtain:

$$\begin{aligned} p_{1|0}pq - p_{1|0}(1-p)(1-q) &\leq p_{1|1}pq - p_{1|1}(1-p)(1-q) \\ &\iff p_{1|0}\big(pq - (1-p)(1-q)\big) \leq p_{1|1}\big(pq - (1-p)(1-q)\big). \end{aligned}$$

This inequality holds true because $p_{1|0} \leq p_{1|1}$ and $\big(pq - (1-p)(1-q)\big) \geq 0$. Therefore, we can conclude that $\Pr(Y = 1 \,|\, Z = 0) \leq \Pr(Y = 1 \,|\, Z = 1)$.

Step 3: From steps 1 and 2, we can conclude that the problem we are interested in becomes the following optimization problem.

$$\begin{aligned} \min_{p,q} \quad & \frac{\dfrac{p_{1|0}p_0(1-p) + p_{1|1}p_1(q)}{p_0(1-p) + p_1 q} - \dfrac{p_{1|0}p_0 p + p_{1|1}p_1(1-q)}{p_0 p + p_1(1-q)}}{p_{1|1} - p_{1|0}} \\ s.t. \quad & \frac{1}{2} \leq p \leq 1, \quad \frac{1}{2} \leq q \leq 1, \\ & p \leq e^{\varepsilon}(1-q), \quad q \leq e^{\varepsilon}(1-p) \end{aligned} \tag{B.0.1}$$

Now, we simplify the objective function of the optimization problem above.

$$\begin{aligned} &\frac{\dfrac{p_{1|0}p_0(1-p) + p_{1|1}p_1(q)}{p_0(1-p) + p_1 q} - \dfrac{p_{1|0}p_0 p + p_{1|1}p_1(1-q)}{p_0 p + p_1(1-q)}}{p_{1|1} - p_{1|0}} \\ &= \frac{p_{1|0}p_0p_1(1-p)(1-q) + p_{1|1}p_0p_1pq - p_{1|0}p_0p_1pq - p_{1|1}p_0p_1(1-p)(1-q)}{\big(p_0(1-p) + p_1 q\big)\big(p_0 p + p_1(1-q)\big)\big(p_{1|1} - p_{1|0}\big)} \end{aligned}$$

$$
\begin{aligned}
&= \frac{\big(p_0 p_1\big)\big(p_{1|0}(1-p)(1-q) + p_{1|1}pq - p_{1|0}pq - p_{1|1}(1-p)(1-q)\big)}{\big(p_0(1-p)+p_1 q\big)\big(p_0 p + p_1(1-q)\big)\big(p_{1|1}-p_{1|0}\big)} \\
&= \frac{\big(p_0 p_1\big)\big(p_{1|1}-p_{1|0}\big)\big(pq-(1-p)(1-q)\big)}{\big(p_0(1-p)+p_1 q\big)\big(p_0 p + p_1(1-q)\big)\big(p_{1|1}-p_{1|0}\big)} \\
&= \frac{\big(p_0 p_1\big)\big(pq-(1-p)(1-q)\big)}{\big(p_0(1-p)+p_1 q\big)\big(p_0 p + p_1(1-q)\big)}
\end{aligned} \tag{B.0.2}
$$

Since the optimization is over the parameters p and q, (B.0.1) is equivalent to the following optimization problem:

$$
\begin{aligned}
\min_{p,q} \quad & \frac{pq-(1-p)(1-q)}{\big(p_0(1-p)+p_1 q\big)\big(p_0 p + p_1(1-q)\big)} \\
s.t. \quad & \frac{1}{2} \le p \le 1, \quad \frac{1}{2} \le q \le 1, \\
& p \le e^{\varepsilon}(1-q), \quad q \le e^{\varepsilon}(1-p)
\end{aligned} \tag{B.0.3}
$$

Suppose ε is given and we want to design the optimal ε-LDP mechanism. p and q are the probability of truly reporting the sensitive attribute of an individual if the original sensitive attribute is 0 and 1 respectively. For a given ε, higher values of p and q yield improved utility. Let $p < e^{\varepsilon}(1-q)$ and $q < e^{\varepsilon}(1-p)$. In this case, for a given ε, we can increase both p and q such that one of the inequalities becomes equality; i.e., we have either $p = e^{\varepsilon}(1-q)$ and $q \le e^{\varepsilon}(1-p)$ or $p \le e^{\varepsilon}(1-q)$ and $q = e^{\varepsilon}(1-p)$. Therefore, we are only interested in those pairs of (p, q) that satisfy one of the two constraints.

Note that these new constraints guarantee non-trivial utility. In other words, for a given ε, we consider the best possible utility (the smallest possible lying probability). This aligns with our initial goal of optimizing the objective function $\min_{\substack{\varepsilon_0-\text{LDP}M \\ \varepsilon_0 \ge \varepsilon}} \frac{\Delta'(D_M^{\varepsilon_0})}{\Delta'(D)}$.

We consider two cases. in case 1 we have: $\big(p = e^{\varepsilon}(1-q)$ and $q \le e^{\varepsilon}(1-p)\big)$ and

in case 2 we have $\big(q = e^{\varepsilon}(1-p)$ and $p \leq e^{\varepsilon}(1-q)\big)$.

Case 1: Here we have $p = e^{\varepsilon}(1-q)$. The objective function becomes:

$$\begin{aligned} \min_{q} \quad & \frac{e^{\varepsilon}(1-q)q - (1-e^{\varepsilon}(1-q))(1-q)}{\big(p_0(1-e^{\varepsilon}(1-q)) + p_1 q\big)\big(p_0 e^{\varepsilon}(1-q) + p_1(1-q)\big)} \\ s.t. \quad & \frac{1}{2} \leq e^{\varepsilon}(1-q) \leq 1, \quad \frac{1}{2} \leq q \leq 1, \\ & q \leq e^{\varepsilon}(1-e^{\varepsilon}(1-q)) \end{aligned}$$

Rearranging the constraints, we can rewrite the objective function of case 1 as:

$$\begin{aligned} \min_{q} \quad & \frac{e^{\varepsilon}(1-q)q - (1-e^{\varepsilon}(1-q))(1-q)}{\big(p_0(1-e^{\varepsilon}(1-q)) + p_1 q\big)\big(p_0 e^{\varepsilon}(1-q) + p_1(1-q)\big)} \\ s.t. \quad & 1 - e^{-\varepsilon} \leq q \leq 1 - \frac{e^{-\varepsilon}}{2}, \quad \frac{1}{2} \leq q \leq 1, \\ & q \geq \frac{e^{\varepsilon}}{e^{\varepsilon}+1} \end{aligned} \tag{B.0.4}$$

Since $1 - e^{-\varepsilon} < \frac{e^{\varepsilon}}{e^{\varepsilon}+1}$, we can further simplify the objective function as:

$$\begin{aligned} \min_{q} \quad & \frac{e^{\varepsilon}(1-q)q - (1-e^{\varepsilon}(1-q))(1-q)}{\big(p_0(1-e^{\varepsilon}(1-q)) + p_1 q\big)\big(p_0 e^{\varepsilon}(1-q) + p_1(1-q)\big)} \\ s.t. \quad & \frac{e^{\varepsilon}}{e^{\varepsilon}+1} \leq q \leq 1 - \frac{e^{-\varepsilon}}{2}, \quad \frac{1}{2} \leq q \leq 1 \end{aligned}$$

Now, we focus on the objective function of the problem (B.0.4).

$$\frac{e^{\varepsilon}(1-q)q - (1-e^{\varepsilon}(1-q))(1-q)}{\big(p_0(1-e^{\varepsilon}(1-q)) + p_1 q\big)\big(p_0 e^{\varepsilon}(1-q) + p_1(1-q)\big)}$$

$$
\begin{aligned}
&= \frac{\big(1-q\big)\big(e^{\varepsilon}q-(1-e^{\varepsilon}(1-q))\big)}{\big(p_0(1-e^{\varepsilon}(1-q))+p_1q\big)\big(p_0e^{\varepsilon}+p_1\big)\big(1-q\big)} \\
&= \frac{\big(e^{\varepsilon}q-(1-e^{\varepsilon}(1-q))\big)}{\big(p_0(1-e^{\varepsilon}(1-q))+p_1q\big)\big(p_0e^{\varepsilon}+p_1\big)} \\
&= \frac{\big(e^{\varepsilon}-1\big)}{\big(p_0(1-e^{\varepsilon}(1-q))+p_1q\big)\big(p_0e^{\varepsilon}+p_1\big)}.
\end{aligned}
$$

Given the constraints that $\frac{e^{\varepsilon}}{e^{\varepsilon}+1} \leq q \leq 1-\frac{e^{-\varepsilon}}{2}$ and $\frac{1}{2} \leq q \leq 1$, since the optimization is over the variable q, we have:

$$
\begin{aligned}
\arg\min_q \quad \frac{\big(e^{\varepsilon}-1\big)}{\big(p_0(1-e^{\varepsilon}(1-q))+p_1q\big)\big(p_0e^{\varepsilon}+p_1\big)} &= \arg\min_q \quad \frac{1}{\big(p_0(1-e^{\varepsilon}(1-q))+p_1q\big)} \\
&= \arg\max_q \quad \big(p_0(1-e^{\varepsilon}(1-q))+p_1q\big) \\
&= \arg\max_q \quad \big(p_0-p_0e^{\varepsilon}+q(p_0e^{\varepsilon}+p_1)\big) \\
&= \arg\max_q \quad q
\end{aligned}
$$

Given the constraints $\frac{e^{\varepsilon}}{e^{\varepsilon}+1} \leq q \leq 1-\frac{e^{-\varepsilon}}{2}$ and $\frac{1}{2} \leq q \leq 1$, the optimal value of q will be $q^* = 1-\frac{e^{-\varepsilon}}{2}$. Therefore, in Case 1, the optimal solution is at $(p,q) = \Big(\frac{1}{2}, 1-\frac{e^{-\varepsilon}}{2}\Big)$. Similarly, in Case 2, the optimal value occurs at $(p,q) = \Big(1-\frac{e^{-\varepsilon}}{2}, \frac{1}{2}\Big)$. Hence, the minimum value arises from either of these cases. Now we want to see when each of the cases is optimal. In order for Case 1 to be the optimal value we should have:

$$
\frac{\frac{1}{2}(1-\frac{e^{-\varepsilon}}{2})-\frac{1}{2}(\frac{e^{-\varepsilon}}{2})}{\big(p_0(\frac{1}{2})+p_1(1-\frac{e^{-\varepsilon}}{2})\big)\big(p_0(\frac{1}{2})+p_1(\frac{e^{-\varepsilon}}{2})\big)} \leq \frac{(1-\frac{e^{-\varepsilon}}{2})(\frac{1}{2})-(\frac{e^{-\varepsilon}}{2})(\frac{1}{2})}{\big(p_0(\frac{e^{-\varepsilon}}{2})+p_1(\frac{1}{2})\big)\big(p_0(1-\frac{e^{-\varepsilon}}{2})+p_1(\frac{1}{2})\big)}
$$

By rearranging, the inequality becomes:

$$\Big(p_0(\frac{e^{-\varepsilon}}{2}) + p_1(\frac{1}{2})\Big)\Big(p_0(1-\frac{e^{-\varepsilon}}{2}) + p_1(\frac{1}{2})\Big) \leq \Big(p_0(\frac{1}{2}) + p_1(1-\frac{e^{-\varepsilon}}{2})\Big)\Big(p_0(\frac{1}{2}) + p_1(\frac{e^{-\varepsilon}}{2})\Big)$$

Therefore, we must have:

$$p_0^2(\frac{e^{-\varepsilon}}{2})(1-\frac{e^{-\varepsilon}}{2}) + p_1^2(\frac{1}{4}) \leq p_1^2(\frac{e^{-\varepsilon}}{2})(1-\frac{e^{-\varepsilon}}{2}) + p_0^2(\frac{1}{4})$$

Or, equivalently:

$$p_1^2 \leq p_0^2 \implies p_1 \leq p_0$$

Specifically, when $p_0 < p_1$, the optimal (p, q) is $\Big(1-\frac{e^{-\varepsilon}}{2}, \frac{1}{2}\Big)$, and when $p_1 < p_0$, it is $(p, q) = \Big(\frac{1}{2}, 1-\frac{e^{-\varepsilon}}{2}\Big)$. If $p_0 = p_1$, then both $\Big(1-\frac{e^{-\varepsilon}}{2}, \frac{1}{2}\Big)$ and $\Big(1-\frac{e^{-\varepsilon}}{2}, \frac{1}{2}\Big)$ will be optimal solutions. We know that in randomized response $p = q = \frac{e^{\varepsilon}}{e^{\varepsilon}+1}$ which is clearly not the optimal solution.

□

Proof of Theorem 15. Let us recall the definitions from the main text that are used in this proof. The definition of $\Delta'(D)$ and $\Delta_{\text{SP}}(\hat{h})$ are as follows:

$$\Delta'(D) = \max_{a,a' \in [k]} \left|\Pr(Y = 1 \mid A = a) - \Pr(Y = 1 \mid A = a')\right|.$$

$$\Delta_{\text{SP}}(\hat{h}) = \max_{a,a' \in [k]} \left|\Pr(\hat{Y} = 1 \mid A = a) - \Pr(\hat{Y} = 1 \mid A = a')\right|.$$

We know that $\Delta'(P_{XAY}) \leq \Delta'(Q_{XAY})$ and we want to prove that $\Delta_{\text{SP}}(h_P) \leq \Delta_{\text{SP}}(h_Q)$.

In order to prove this theorem, we refer to Theorem 1 in Kamiran and Calders (2012).

Theorem 1: A classifier h is DA-optimal in $\mathcal{H}^*$ iff

$$\text{acc}\left(h^{\text{Perf}}\right) - \text{acc}(h) = \frac{2n_0n_1}{(n_0+n_1)^2}\left(\Delta_{\text{SP}}\left(h^{\text{Perf}}\right) - \Delta_{\text{SP}}(h)\right),$$

where $\mathcal{H}^*$ denotes the class of all classifiers satisfying $\Pr(\hat{Y}=1) = \Pr(Y=1)$. h^{Perf} refers to a perfect classifier, and n_0 and n_1 denote number of data points with $A=0$ and $A=1$ on a dataset used for training respectively.

Since $\Pr_{(X,A,Y)\sim P_{XAY}}(A=a) = \Pr_{(X,A,Y)\sim Q_{XAY}}(A=a)$ for $a \in \{0,1\}$, it follows that $\frac{2n_0n_1}{(n_0+n_1)^2}$ is a fixed value for two distributions P_{XAY} and Q_{XAY}. We denote this constant by c. In addition, since h_P and h_Q are DA-optimal classifiers in $\mathcal{H}^*$, we have:

$$1 - \text{acc}(h_P) = c\left(\Delta'\left(P_{XAY}\right) - \Delta_{\text{SP}}(h_P)\right)$$
$$1 - \text{acc}(h_Q) = c\left(\Delta'\left(Q_{XAY}\right) - \Delta_{\text{SP}}(h_Q)\right).$$

Note that we used $\Delta'\left(P_{XAY}\right)$ and $\Delta'\left(Q_{XAY}\right)$ rather than $\Delta_{\text{SP}}\left(h^{\text{Perf}}\right)$ since the perfect classifier identically represents the unfairness of the data distribution.

Since we have assumed that the accuracy of learned classifiers h_P and h_Q are equal, we have:

$$\left(\Delta'\left(P_{XAY}\right) - \Delta_{\text{SP}}(h_P)\right) = \left(\Delta'\left(Q_{XAY}\right) - \Delta_{\text{SP}}(h_Q)\right).$$

From this equation, we can conclude that if $\Delta'\left(P_{XAY}\right) \leq \Delta'\left(Q_{XAY}\right)$, then $\Delta_{\text{SP}}(h_P) \leq \Delta_{\text{SP}}(h_Q)$. □

Lemma 16. *$\Delta(D) \leq c_1\Delta'(D)$ and $\Delta'(D) \leq c_2\Delta(D)$ for constants c_1 and c_2 dependent on the marginal distribution P_Y of the joint distribution P_{XAY}.*

For the first part, we have:

Proof.

$$\begin{aligned}
\Delta(D) &= \max_{a\in[k]} \left| \frac{\Pr(Y=1 \mid A=a)}{\Pr(Y=1)} - 1 \right| \\
&= \max_{a\in[k]} \left| \frac{\Pr(Y=1 \mid A=a) - \Pr(Y=1)}{\Pr(Y=1)} \right| \\
&\leq \max_{a,a'\in[k]} \left| \frac{\Pr(Y=1 \mid A=a) - \Pr(Y=1 \mid A=a')}{\Pr(Y=1)} \right| \\
&= \frac{1}{\Pr(Y=1)} \Delta'(D),
\end{aligned}$$

where the third line follows from the fact that $\min\limits_{a'\in[k]} |\Pr(Y=1 \mid A=a')| \leq \Pr(Y=1)$.

Additionally, for the second part, it can be shown that:

$$\begin{aligned}
\Delta'(D) &= \max_{a,a'\in[k]} \Big| \Pr(Y=1 \mid A=a) - \Pr(Y=1 \mid A=a') \Big| \\
&= \max_{a,a'\in[k]} \Big| \Pr(Y=1 \mid A=a) - \Pr(Y=1) + \Pr(Y=1) - \Pr(Y=1 \mid A=a') \Big| \\
&\leq \max_{a,a'\in[k]} \left[\Big| \Pr(Y=1 \mid A=a) - \Pr(Y=1) \Big| + \Big| \Pr(Y=1 \mid A=a') - \Pr(Y=1) \Big| \right] \\
&\leq 2 \max_{a\in[k]} \Big| \Pr(Y=1 \mid A=a) - \Pr(Y=1) \Big| \\
&= 2 \Pr(Y=1) \Delta(D).
\end{aligned}$$

It follows that $c_1 = \frac{1}{\Pr(Y=1)}$ and $c_2 = 2\Pr(Y=1)$, where c_1 and c_2 only depend on the marginal distribution P_Y of the joint distribution P_{XAY}. □

Bibliography

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR.

Agarwal, S. (2020). Trade-offs between fairness, interpretability, and privacy in machine learning. UWSpace. http://hdl.handle.net/10012/15861.

Alghamdi, W., Hsu, H., Jeong, H., Wang, H., Michalak, P., Asoodeh, S., and Calmon, F. (2022). Beyond adult and compas: Fair multi-class prediction via information projection. *Advances in Neural Information Processing Systems*, **35**, 38747–38760.

Arcolezi, H. H., Makhlouf, K., and Palamidessi, C. (2023). (local) differential privacy has no disparate impact on fairness. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 3–21. Springer.

Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. (2019). Differential privacy has

disparate impact on model accuracy. *Advances in neural information processing systems*, **32**.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, **50**(1), 3–44.

Calders, T., Kamiran, F., and Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops*, pages 13–18. IEEE.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, **30**.

Celis, L. E., Keswani, V., and Vishnoi, N. (2020). Data preprocessing to mitigate bias: A maximum entropy based approach. In *International conference on machine learning*, pages 1349–1359. PMLR.

Chakraborty, J., Majumder, S., and Menzies, T. (2021). Bias in machine learning software: Why? how? what to do? In *Proceedings of the 29th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*, pages 429–440.

Chang, H. and Shokri, R. (2021a). On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 292–303. IEEE.

Chang, H. and Shokri, R. (2021b). On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 292–303. IEEE.

Chen, C., Liang, Y., Xu, X., Xie, S., Kundu, A., Payani, A., Hong, Y., and Shu, K. (2022). When fairness meets privacy: Fair classification with semi-private sensitive attributes. *arXiv preprint arXiv:2207.08336*.

Cho, J., Hwang, G., and Suh, C. (2020a). A fair classifier using kernel density estimation. *Advances in neural information processing systems*, **33**, 15088–15099.

Cho, J., Hwang, G., and Suh, C. (2020b). A fair classifier using mutual information. In *2020 IEEE international symposium on information theory (ISIT)*, pages 2521–2526. IEEE.

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

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. (2019). Leveraging labeled and unlabeled data for consistent fair binary classification. *Advances in Neural Information Processing Systems*, **32**.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806.

Cummings, R., Gupta, V., Kimpara, D., and Morgenstern, J. (2019). On the compatibility of privacy and fairness. In *Adjunct publication of the 27th conference on user modeling, adaptation and personalization*, pages 309–315.

Denis, C., Elie, R., Hebiri, M., and Hu, F. (2024). Fairness guarantees in multi-class classification with demographic parity. *Journal of Machine Learning Research*, **25**(130), 1–46.

Differential privacy team Apple (2017). Learning with privacy at scale.

Ding, J., Zhang, X., Li, X., Wang, J., Yu, R., and Pan, M. (2020). Differentially private and fair classification via calibrated functional mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 622–629.

Dwork, C. (2006). Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Dwork, C., Roth, A., *et al.* (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, **9**(3–4), 211–407.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. (2018). Decoupled classifiers for group-fair and efficient machine learning. In *Conference on fairness, accountability and transparency*, pages 119–133. PMLR.

Erlingsson, Ú., Pihur, V., and Korolova, A. (2014). Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067.

Esipova, M. S., Ghomi, A. A., Luo, Y., and Cresswell, J. C. (2023). Disparate impact in differential privacy from gradient misalignment. In *The Eleventh International Conference on Learning Representations*.

Evfimievski, A., Gehrke, J., and Srikant, R. (2003). Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 211–222.

Farrand, T., Mireshghallah, F., Singh, S., and Trask, A. (2020). Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In *Proceedings of the 2020 workshop on privacy-preserving machine learning in practice*, pages 15–19.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268.

Ganev, G., Oprisanu, B., and De Cristofaro, E. (2022). Robin hood and matthew effects: Differential privacy has disparate impact on synthetic data. In *International Conference on Machine Learning*, pages 6944–6959. PMLR.

Ghoukasian, H. and Asoodeh, S. (2024). Differentially private fair binary classifications. *arXiv preprint arXiv:2402.15603*.

Gohar, U., Biswas, S., and Rajan, H. (2023). Towards understanding fairness and its composition in ensemble machine learning. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pages 1533–1545. IEEE.

Gopi, S., Lee, Y. T., and Wutschitz, L. (2021). Numerical composition of differential privacy. *Advances in Neural Information Processing Systems*, **34**, 11631–11642.

Hajian, S. and Domingo-Ferrer, J. (2012). A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering*, **25**(7), 1445–1459.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, **29**.

Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Sharifi-Malvajerdi, S., and Ullman, J. (2019). Differentially private fair learning. In *International Conference on Machine Learning*, pages 3000–3008. PMLR.

Jiang, H. and Nachum, O. (2020). Identifying and correcting label bias in machine learning. In *International conference on artificial intelligence and statistics*, pages 702–712. PMLR.

Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. (2020). Wasserstein fair classification. In *Uncertainty in artificial intelligence*, pages 862–872. PMLR.

Jiao, H. and Li, B. (2022). Solving min–max linear fractional programs based on image space branch-and-bound scheme. *Chaos, Solitons & Fractals*, **164**, 112682.

Kairouz, P., Oh, S., and Viswanath, P. (2014). Extremal mechanisms for local differential privacy. *Advances in neural information processing systems*, **27**.

Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, **33**(1), 1–33.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, **30**.

Kifer, D., Messing, S., Roth, A., Thakurta, A., and Zhang, D. (2020). Guidelines for implementing and auditing differentially private systems. *ArXiv*, **abs/2002.04049**.

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, **30**.

Kim, J. S., Chen, J., and Talwalkar, A. (2020). Fact: A diagnostic for group fairness trade-offs. In *International Conference on Machine Learning*, pages 5264–5274. PMLR.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Ko, J., Ziani, J., Das, S., Williams, M., and Fioretto, F. (2024). Fairness issues and mitigations in (differentially private) socio-demographic data processes. *arXiv preprint arXiv:2408.08471*.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. *Advances in neural information processing systems*, **30**.

Lichman, M. (2013). Uci machine learning repository.

Lowy, A., Baharlouei, S., Pavan, R., Razaviyayn, M., and Beirami, A. (2021). A stochastic optimization framework for fair risk minimization. *arXiv preprint arXiv:2102.12586*.

Lowy, A., Gupta, D., and Razaviyayn, M. (2023). Stochastic differentially private and fair learning. In *International Conference on Learning Representations*.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2018). Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR.

Makhlouf, K., Arcolezi, H. H., Zhioua, S., Brahim, G. B., and Palamidessi, C. (2024a). On the impact of multi-dimensional local differential privacy on fairness. *Data Mining and Knowledge Discovery*, pages 1–24.

Makhlouf, K., Stefanović, T., Arcolezi, H. H., and Palamidessi, C. (2024b). A systematic and formal study of the impact of local differential privacy on fairness: Preliminary results. In *2024 IEEE 37th Computer Security Foundations Symposium (CSF)*, pages 1–16. IEEE.

Mangold, P., Perrot, M., Bellet, A., and Tommasi, M. (2023). Differential privacy has bounded impact on fairness in classification. In *International Conference on Machine Learning*, pages 23681–23705. PMLR.

Mary, J., Calauzenes, C., and El Karoui, N. (2019). Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, pages 4382–4391. PMLR.

Mozannar, H., Ohannessian, M., and Srebro, N. (2020). Fair learning with private demographic data. In *International Conference on Machine Learning*, pages 7066–7075. PMLR.

Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., and Erlingsson, Ú. (2018). Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*.

Peng, K., Chakraborty, J., and Menzies, T. (2022). Fairmask: Better fairness via model-based rebalancing of protected attributes. *IEEE Transactions on Software Engineering*, **49**(4), 2426–2439.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. *Advances in neural information processing systems*, **30**.

Prost, F., Qian, H., Chen, Q., Chi, E. H., Chen, J., and Beutel, A. (2019). Toward a better trade-off between performance and fairness with kernel-based distribution matching. *arXiv preprint arXiv:1910.11779*.

Pujol, D., McKenna, R., Kuppam, S., Hay, M., Machanavajjhala, A., and Miklau, G. (2020). Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 189–199.

Rogers, R., Subramaniam, S., Peng, S., Durfee, D., Lee, S., Kancha, S. K., Sahay, S., and Ahammad, P. (2020). Linkedin's audience engagements api: A privacy preserving data analytics system at scale. *arXiv preprint arXiv:2002.05839*.

Rosenblatt, L., Stoyanovich, J., and Musco, C. (2023). A simple and practical method for reducing the disparate impact of differential privacy. *arXiv preprint arXiv:2312.11712*.

Sabato, S., Treister, E., and Yom-Tov, E. (2024). Fairness and unfairness in binary and multiclass classification: Quantifying, calculating, and bounding.

Tran, C., Fioretto, F., and Van Hentenryck, P. (2021a). Differentially private and fair deep learning: A lagrangian dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9932–9939.

Tran, C., Dinh, M., and Fioretto, F. (2021b). Differentially private empirical risk minimization under the fairness lens. *Advances in Neural Information Processing Systems*, **34**, 27555–27565.

Tran, C., Zhu, K., Fioretto, F., and Van Hentenryck, P. (2022). Sf-pate: scalable, fair, and private aggregation of teacher ensembles. *arXiv preprint arXiv:2204.05157*.

Wang, H., Ustun, B., and Calmon, F. (2019). Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In *International Conference on Machine Learning*, pages 6618–6627. PMLR.

Wang, H., Hsu, H., Diaz, M., and Calmon, F. P. (2021). To split or not to split: The impact of disparate treatment in classification. *IEEE Transactions on Information Theory*, **67**(10), 6733–6757.

Wang, H., He, L., Gao, R., and Calmon, F. (2024). Aleatoric and epistemic discrimination: Fundamental limits of fairness interventions. *Advances in Neural Information Processing Systems*, **36**.

Wang, S., Huang, L., Wang, P., Nie, Y., Xu, H., Yang, W., Li, X.-Y., and Qiao, C. (2016). Mutual information optimally local private discrete distribution estimation. *arXiv preprint arXiv:1607.08025*.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American statistical association*, **60**(309), 63–69.

Wei, D., Ramamurthy, K. N., and Calmon, F. P. (2020). Optimized score transformation for fair classification. *Proceedings of Machine Learning Research*, **108**.

Wei, D., Ramamurthy, K. N., and Calmon, F. P. (2021). Optimized score transformation for consistent fair classification. *Journal of Machine Learning Research*, **22**(258), 1–78.

Wightman, L. F. (1998). Lsac national longitudinal bar passage study. lsac research report series.

Xu, D., Yuan, S., and Wu, X. (2019). Achieving differential privacy and fairness in logistic regression. In *Companion proceedings of The 2019 world wide web conference*, pages 594–599.

Xu, D., Du, W., and Wu, X. (2021). Removing disparate impact on model accuracy in differentially private stochastic gradient descent. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1924–1932.

Yaghini, M., Liu, P., Boenisch, F., and Papernot, N. (2023). Learning with impartiality to walk on the pareto frontier of fairness, privacy, and utility. *arXiv preprint arXiv:2302.09183*.

Yang, F., Cisse, M., and Koyejo, S. (2020). Fairness with overlapping groups; a probabilistic perspective. *Advances in neural information processing systems*, **33**, 4067–4078.

Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J., Cormode, G., and Mironov, I. (2021). Opacus: User-friendly differential privacy library in pytorch.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pages 962–970. PMLR.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Zhang, J., Zhang, Z., Xiao, X., Yang, Y., and Winslett, M. (2012). Functional mechanism: Regression analysis under differential privacy.

Zhao, H. and Gordon, G. J. (2022). Inherent tradeoffs in learning fair representations. *The Journal of Machine Learning Research*, **23**(1), 2527–2552.