Socially Responsible Machine Learning: On the Preservation of Individual Privacy and Fairness

by

Xueru Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical and Computer Engineering)
in the University of Michigan
2021

Doctoral Committee:

Mingyan Liu, Chair
Yiling Chen, Harvard University
Alfred Hero
Atul Prakash
Aaron Roth, University of Pennsylvania

Xueru Zhang

xueru@umich.edu

ORCID iD: 0000-0002-0761-5943

To my parents.

ACKNOWLEDGMENTS

I have been lucky to complete this dissertation with the help and support of amazing mentors, colleagues, collaborators, and friends. I want to start by thanking my advisor, Mingyan Liu, for her insights and encouragements, for sharing with me her knowledge of a wide range of research, and for giving me tremendous support. Without her, I wouldn't be able to grow as a researcher. I appreciate all the help and suggestions she provided during my job applications. It has been a true privilege to have her as my advisor.

I am extremely grateful to Professor Alfred Hero, Professor Yang Liu, Dr. Tulga Ersal, and Professor Parinaz Naghizadeh for my references and for supporting my job applications. I would like to express my deepest appreciation to my committee members, Professor Yiling Chen, Professor Aaron Roth, Professor Alfred Hero, and Professor Atul Prakash, for the time they devoted, the helpful discussions, and insightful feedback.

During my time at the University of Michigan, I have had the privilege to work with many wonderful researchers: Mahdi Khalili, Chunan Huang, Ruibo Tu, Kun Jin, Yang Liu, Parinaz Naghizadeh, Tulga Ersal, Cem Tekin, Anna Stefanopoulou. Thank you for your insights, and I have learned a lot from you.

I am also grateful to my colleagues: Mahdi Khalili, Chaowei Xiao, Chenlan Wang, Kun Jin, Armin Sarabi, Mehrdad Moharrami, Tongxin Yin, Ranjan Pal, Demba Komma. Thank you for your insights and help. It has been wonderful to meet and work with you. Special thanks to Mahdi, thank you for being a constant source of support and encouragement. Chenlan, thank you for helping and accompanying me all these years in Ann Arbor.

Lastly, I want to thank my parents for their endless love and support. Thank you for having my back and always believing in me. None of my achievements would be possible without you. I also want to express my gratitude to my grandparents, aunts, uncles, and cousins. I love you!

This work is supported by the NSF under grants CNS-1616575, CNS-1646019, CNS-1739517, CNS-2040800, and by the ARO under contract W911NF1810208.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Machine learning (ML) techniques have seen significant advances over the last decade and are playing an increasingly critical role in people's lives. While their potential societal benefits are enormous, they can also inflict great harm if not developed or used with care. In this thesis, we focus on two critical ethical issues in ML systems, the violation of privacy and fairness, and explore mitigating approaches in various scenarios.

On the privacy front, when ML systems are developed with private data from individuals, it is critical to prevent privacy violation. Differential privacy (DP), a widely used notion of privacy, ensures that no one by observing the computational outcome can infer a particular individual's data with high confidence. However, DP is typically achieved by randomizing algorithms (e.g., adding noise), which inevitably leads to a trade-off between individual privacy and outcome accuracy. This trade-off can be difficult to balance, especially in settings where the same or correlated data is repeatedly used/exposed during the computation. In the first part of the thesis, we illustrate two key ideas that can be used to balance an algorithm's privacy-accuracy tradeoff: (1) the reuse of intermediate computational results to reduce information leakage; and (2) improving algorithmic robustness to accommodate more randomness. We introduce a number of randomized, privacy-preserving algorithms that leverage these ideas in various contexts such as distributed optimization and sequential computation. It is shown that our algorithms can significantly improve the privacy-accuracy tradeoff over existing solutions.

On the fairness front, ML systems trained with real-world data can inherit biases and exhibit discrimination against already-disadvantaged or marginalized social groups. Recent works have proposed many fairness notions to measure and remedy such biases. However, their effectiveness is mostly studied in a static framework without accounting for the interactions between individuals and ML systems. Since individuals inevitably react to the algorithmic decisions they are subjected to, understanding the downstream impacts of ML decisions is critical to ensure that these decisions

are socially responsible. In the second part of the thesis, we present our research on evaluating the long-term impacts of (fair) ML decisions. Specifically, we establish a number of theoretically rigorous frameworks to model the interactions and feedback between ML systems and individuals, and conduct equilibrium analysis to evaluate the impact they each have on the other. We will illustrate how ML decisions and individual behavior evolve in such a system, and how imposing common fairness criteria intended to promote fairness may nevertheless lead to undesirable pernicious effects. Aided with such understanding, mitigation approaches are also discussed.

CHAPTER 1

Introduction

The development of Machine learning (ML) techniques has revolutionized people's daily lives and enabled breakthroughs in various scientific fields such as robotics, computer vision, natural language processing, etc. Despite the enormous societal benefits, ML techniques have also caused ethical concerns when used to make consequential decisions about humans. It has become evident that in many such domains (e.g., lending, hiring, criminal justice, healthcare, etc.), ML techniques can behave in unintended and potentially harmful ways: (1) they may expose individuals who have contributed their data to risks; (2) they may result in adverse outcomes for people who are affected by decisions made by ML algorithms. In this thesis, we will focus on two critical issues that arise from these two types of issues: (i) *privacy violation*; and (ii) *discrimination*.

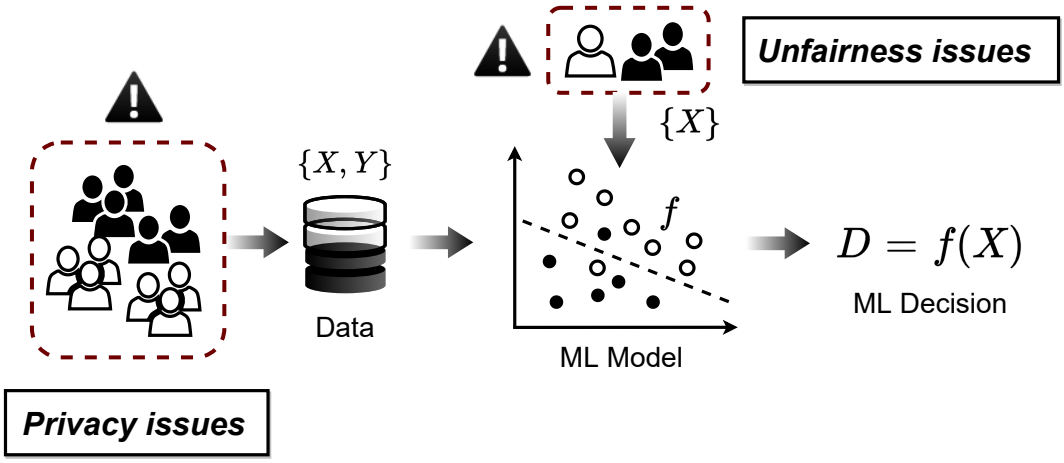


Figure 1.1: Misuse of ML techniques

Privacy violation. When ML algorithms are developed using individuals' data such as medical records, financial data, online activities, etc., their privacy is at high risk of being compromised, resulting in potentially significant harm and monetary losses to both data collectors and data owners. For example, the Federal Trade Commission (FTC) in the U.S. has levied a $5 billion penalty against Facebook and $170 million penalty against Google for violating consumer privacy in 2019 [2]. During 2019, FTC's Consumer Sentinel Network has received nearly $1.7 million fraud reports and consumers have lost more than $1.9 billion to frauds in total [3]. In addition to losses, privacy concerns have become a major source of distrust and a major obstacle to people sharing their personal data with data analysts, resulting in a lack of sufficient data to develop robust and accurate computational models. Therefore, understanding how privacy violations happen and developing solutions to address this issue are very important.

There are many reasons for privacy violations. One of the most direct causes is excessive data collection and sharing. In this era of big data, user/consumer data has been over-collected by many online platforms, mobile apps, and third-party trackers. A recent study [15] shows that among 959,000 Android apps, nearly 90% of them were set up to transfer collected personal information back to Google. There is a lack of transparency in data collection, usage, and sharing. The emergence of data regulation such as the European Union's General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA) aims to impose constraints on data collection and transfer, but they are likely insufficient. Even without access to data, personal information may be inferred from outputs of ML models. A prime example is the model inversion attacks studied in [46], where a facial recognition API is trained based on a set of face images. An attacker has no access to the model parameters of the API but can send images to the API. For each image the attacker sends, it receives a vector of confidence scores along with the name of the face as recognized from the input image. [46] shows that the attacker, using the outputs from sending a set of randomly generated face images, can reconstruct the facial appearance of a person who has contributed to the training dataset.

Some studies address privacy and security issues from game theoretical perspectives [84–87, 89–94]. Another effective approach to preventing privacy violation is to process sensitive data with privacy-preserving algorithms. The first step is defining individual privacy and understanding what it means to protect individual privacy in data analysis. One class of privacy notions, including k-anonymity [128], l-diversity [107], t-closeness [98], is based on anonymization, where the idea is to de-identify the dataset by removing personal identifiable information (e.g., name, SSN, gender)

and release the anonymous version of the dataset. However, these anonymization-based approaches are not resilient to "linkage" attacks. One notable example is the Netflix prize competition, where the company published its user-movie rating dataset with all users' names removed. From this anonymous dataset, researchers can design new movie recommendation systems. However, this anonymous dataset turned out to be far from private: most of the users in the dataset can be de-anonymized by cross-linking the film ratings on the public Internet Movie Database (IMDb) [115]. This example also shows the need for a mathematically rigorous notion of privacy that is suitable for complex data analysis and resilient to any attackers regardless of their background knowledge. *Differential privacy (DP)* [35], as a widely used notion of privacy satisfying these requirements, ensures that no one by observing the computational outcome can infer with substantial higher confidence than random guessing whether a particular individual's data was included in the data analysis or not. However, DP guarantee is typically achieved by randomizing algorithms (e.g., adding noise), which inevitably leads to the tradeoff between individual privacy and the accuracy of outcomes. This tradeoff can be difficult to balance, especially for settings where the same or correlated data is used multiple times over the computational process. Because individual privacy leakage accumulates substantially (as more information about the same data is revealed), controlling total privacy leakage with sufficient accuracy becomes particularly challenging and important when designing the private algorithms.

Discrimination. Over the last decade, an increasing number of ML algorithms have been developed to help make high-stakes decisions about real people; these include domains such as hiring (e.g., HireVue, Gild, Entelo), lending (e.g., Wonga), criminal justice (e.g., COMPAS, PredPol) to name a few. On the one hand, these ML models can find hidden patterns and intrinsic structures in the input data with high accuracy. On the other hand, these models can inherit pre-existing bias in the dataset and exhibit discrimination against protected population groups. It has been well-documented that ML algorithms can exhibit or even exacerbate such discrimination in many real-world applications. For example, a study shows that speech recognition products such as Amazon's Alexa and Google Home can have accent bias, with Chinese-accented and Spanish-accented English hardest to understand [58]. The COMPAS recidivism prediction tool, used by courts in the US in parole decisions, has been shown to have a substantially higher false positive rate for African Americans compared to the general population [28]. Amazon had been using automated software since 2014 to assess applicants' resumes, which were found to be biased against women [30].

There are various potential causes for such discrimination. It may have been introduced when data is collected. For instance, if data sampled from a minority group is much smaller in size than that from a majority group, then the model could be more in favor of the majority group due to this representation disparity (e.g., more than a third of data in ImageNet and Open Images, two datasets widely used in machine learning research communities, is US-based [123]). Another example is when the data collection decision itself reflects bias, which then impacts the collected data (e.g., if more police officers are dispatched to places believed to have higher crime rate to begin with, then crimes are more likely to be recorded in these places [38]). Even when the data collection process is unbiased, bias may already exist in the data. Historical prejudice and stereotypes can be preserved in data (e.g., the relationship between "man" and "computer programmers" were found to be similar to that between "woman" and "homemaker" [17]).

One commonly used approach to alleviating the discrimination issue is to enforce certain fairness constraints upon the training process. Depending on the applications, a variety of fairness constraints have been proposed and can be roughly classified into two families: (1) *group fairness* aims to achieve a certain balance in group-level: the whole population is partitioned into a small number of protected groups distinguished based on some sensitive attributes (e.g., race, gender), and it requires certain statistical measure (e.g., positive rates, true positive rates, etc.) to be approximately equalized across different protected groups; (2) *individual fairness* is in pursuit of equity in individual level: it requires that similar individuals be treated similarly.

While the effectiveness of these fairness constraints has been shown in various domains, most of the studies are done under a static framework where only the immediate impact of the constraints is assessed but not its long-term consequences. Because ML models are deployed in a dynamic environment, people may change their behaviors in response to the perceived decisions, and such change can further be captured in future models [144, 152]. Under this complex interplay between algorithmic decisions and individuals' reactions, the fairness criteria that intend to protect disadvantaged groups may lead to unintended, pernicious long-term effects [103]. Consider an example in lending where a lender decides whether or not to issue a loan based on the applicant's credit score. Decisions satisfying an identical true positive rate across different racial groups can make the outcome seem fairer [57]. However, this can potentially result in more loans issued to less qualified applicants in the group whose score distribution skews toward higher default risk. The lower repayment among these individuals causes their future credit scores to drop, which moves the group's score distribution further toward higher default risk [103]. Therefore, understanding how

algorithmic decisions and people interact over time and examining the long-term impact of fairness criteria is essential when developing fair ML systems, which can be challenging due to the lack of dynamic datasets and models that characterize the human behaviors.

1.1 Background

Before discussing the contributions of this thesis in more depth, we present notions of privacy and fairness that we used and studied.

1.1.1 Differential Privacy

Differential privacy, first proposed by Dwork, McSherry, Nissim, and Smith [35], centers around the idea that the output of a certain mechanism or computational procedure should be statistically similar given singular changes to the input, thereby preventing meaningful inference from observing the output.

To illustrate the guarantee of differential privacy, consider an attacker aiming at inferring private information of a target individual, whose data may or may not be contributed to the dataset in a computation. The attacker is able to observe the computational outcome, and may have access to any arbitrary side information (e.g., the private data of every other individual in the dataset, some knowledge about the target individual, etc.). Differential privacy guarantees that regardless of what side information the attacker has, the attacker can learn almost nothing new about the target individual from the computational outcome.

Formally, a randomized algorithm $\mathcal{A}(\cdot)$ taking dataset $D \in \mathcal{D}$ as input satisfies (ϵ, δ)-differential privacy if for any datasets $D, \widehat{D}$ that are different in at most one individual's data and for any set of possible outputs $\mathcal{S} \in \text{range}(\mathcal{A})$, we have

$$\Pr(\mathcal{A}(D) \in \mathcal{S}) \leq \Pr(\mathcal{A}(\widehat{D}) \in \mathcal{S}) + \delta,$$

where $\epsilon \geq 0$ bounds the privacy loss, and $\delta \in [0, 1]$ loosely corresponds to the probability that the algorithm fails to bound the privacy loss by ϵ. In particular, when $\delta = 0$ we omit it and say algorithm $\mathcal{A}$ preserves ϵ-differential privacy.

Differential privacy is a worst case measure, that is, the bound is over all possible random outputs and all possible inputs. It is a strong guarantee, as it can protect against attackers with any

side information. Moreover, it admits a powerful algorithmic framework. There are two important properties that make differential privacy easily be used for complex data analysis. The first is *immunity to post-processing*: a differentially private output followed by any data-independent computation remains satisfying differential privacy. The second is *composability*: when a differentially private algorithm is queried independently over the same data multiple times, the total privacy loss accumulates (i.e., privacy guarantee degrades).

1.1.2 Fairness in Supervised Learning

In supervised learning, the goal is to predict a true outcome Y from features X based on labeled training data. To ensure the prediction $\widehat{Y}$ is non-discriminatory, certain fairness criterion should be satisfied. As mentioned in introduction, a variety of fairness criteria have been formulated to measure and remedy biases in machine learning systems. In this thesis, we focus on group fairness where the population is partitioned into a small number of groups distinguished by some sensitive attributes $S \in \mathcal{S}$, and certain statistics are equalized across different groups. In particular, we are interested in studying two criteria called *demographic parity* and *equal opportunity*.

1. *Demographic Parity (`DP`)* [11]: prediction $\widehat{Y}$ is independent of group sensitive attribute S.
2. *Equal Opportunity (`EqOpt`)* [57]: prediction $\widehat{Y}$ is conditional independent of group sensitive attribute S given true outcome Y.

For binary classification where $Y \in \{0,1\}, \widehat{Y} \in \{0,1\}$, `DP` requires the positive classification rates to be equalized across different groups, i.e., $\Pr(\widehat{Y}=1|S=s) = \Pr(\widehat{Y}=1), \forall s \in \mathcal{S}$, while `EqOpt` requires true positive rates to be equalized across different groups, i.e., $\Pr(\widehat{Y}=1|Y=1,S=s) = \Pr(\widehat{Y}=1|Y=1), \forall s \in \mathcal{S}$. To interpret these two criteria, consider settings such as hiring, lending, and college admissions, where a decision maker (e.g., company, bank, college) aims to select ($\widehat{Y}=1$) individuals from the applicant pool that are qualified ($Y=1$) for certain tasks based on a given set of features X. `DP` fairness ensures that all applicants from different groups are selected at the same rate, while `EqOpt` fairness is only concerned with the equity among the qualified individuals and it ensures that the qualified applicants from different groups are selected at the same rate.

A variety of methods has been proposed for learning fair supervised learning models, and they can be roughly classified into three families,

1. *Pre-processing:* remove pre-existing biases by modifying the datasets before training process [78, 143].

2. *In-processing:* impose fairness constraint during the training process, typically by adding a constraint to optimization problem or changing objective function [5, 141, 142].
3. *Post-processing:* adjust the output of an algorithm based on group sensitive attributes after training process [57, 120].

1.2 Overview of Thesis Contributions and Structure

Part I: Designing Differentially Private Algorithms

One approach to leveraging individuals' data while preventing privacy violation, is to process sensitive data with privacy-preserving algorithms. In the first part of this dissertation, we present a number of differentially private algorithms for multiple computational tasks including *distributed learning* (Chapter 2) and *sequential computations* (Chapter 3).

During these computations, the same or correlated data is repeatedly used/exposed during these computations. Specifically, in distributed learning, multiple entities collaboratively work through an interactive process of local computation (over local, private data) and message passing; during this interactive process the same local data is repeatedly used. In sequential computations, individual's temporal data is generated/acquired sequentially for online analysis, and there is the strong temporal correlation within the data sequence. Because the same or correlated data is repeatedly used, the total privacy leakage of each individual accumulates substantially over time during the computation. As such, balancing the trade-off between individual privacy and the outcome accuracy can be challenging.

To improve the privacy-accuracy trade-off, we have explored two ideas:

(a) Reuse intermediate computational results to reduce the total information leakage.
(b) Improve algorithmic robustness to accommodate more randomness.

Intuitively, when less information is revealed, less randomization is required to achieve the same privacy guarantee, so that the accuracy can be increased; when an algorithm is more robust, it can accommodate more randomization to enhance privacy without jeopardizing too much accuracy. Based on these ideas, we designed multiple novel algorithms whose privacy-accuracy trade-off is improved significantly over the existing algorithms.

Chapter 2: Private ADMM-Based Distributed Algorithms. In this chapter, we consider a consensus problem in a fully distributed setting where multiple entities collaboratively work toward a common optimization objective through an interactive process of local computation (over local, private data) and message passing. We focused on the Alternating Direction Method of Multiplier (ADMM)-based algorithms to solve the distributed optimization. Because local computations are exchanged among different entities, privacy concerns arise. A differentially private ADMM was proposed in prior work [147] where only the privacy loss of a single node during one iteration was bounded, a method that makes it difficult to balance the tradeoff between the utility attained through distributed computation and privacy guarantees when considering the total privacy loss of all nodes over the entire iterative process. To improve privacy-accuracy trade-off, we leverage idea (a),(b) and propose a number of algorithms by modifying the original ADMM algorithms. Specifically, R-ADMM [150] utilizes (a) and ensures the privacy leakage only happens in half of the updates; M-ADMM [149] utilizes (b) which improves the algorithmic robustness; MR-ADMM [151] incorporates both ideas to improve the trade-off further.

Chapter 3: Real-Time Release of Sequential Data with Differential Privacy. Many data analytics applications rely on temporal data, generated (and possibly acquired) sequentially for online analysis. In this chapter, we propose an algorithm to release the sequential data in real-time with differential privacy guarantee. Because of the (potentially strong) temporal correlation within the data sequence, the overall privacy loss can accumulate significantly over time; an attacker with statistical knowledge of the correlation can be particularly hard to defend against. An idea that has been explored in the literature to alleviate this problem is to factor this correlation into the perturbation/noise mechanism. Existing work, however, either focuses on the offline setting (where perturbation is designed and introduced after the entire sequence has become available) [44, 80, 121, 133, 138], or requires a priori information on the correlation in generating perturbation [41]. In contrast, the algorithm we propose can learn the correlation as the sequence is generated, and the learned correlation is used for estimating future data in the sequence. This estimate then drives the generation of the noisy released data. This method allows us to design better perturbation and is suitable for real-time operations. We show theoretically and empirically that this approach achieves high accuracy with lower privacy loss compared to existing methods. Furthermore, this method has been used to enable private vehicle-to-vehicle communication in intelligent transportation systems [67, 148].

Part II: Fair Machine Learning with Human in Loops

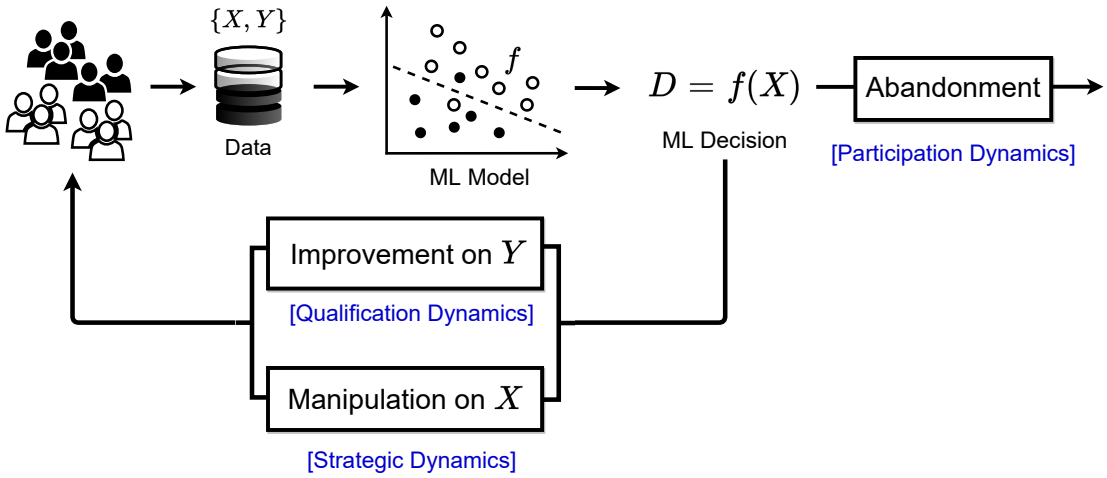


Figure 1.2: ML with human in feedback loops: three types of interactions

Algorithmic fairness in machine learning has been studied extensively in static settings where one-shot (fair) ML decisions are made on tasks such as classification. However, in practice ML models are deployed in a dynamic environment where ML models and individuals feed and affect each other. Without accounting for individuals' behaviors and the interactions, ML decisions and fairness interventions may result in unintended consequences. In the second part of this dissertation, we study the impacts of (fair) ML on the well-being of different social groups. Specifically, We establish multiple theoretically rigorous frameworks to model the interactions between ML systems and individuals, and conduct equilibrium analysis to evaluate the impacts they each have on each other. Depending on how individuals respond to the perceived decisions, we consider the following three types of interactions,

1. *Participation Dynamics* (Chapter 4)
   ML decisions made in the past may affect individuals' participation/retention in the future ML systems. Consider an example of speech recognition, individuals are more likely to leave the system if they experience the lower accuracy; this in turn can affect the group representation in future datasets used for building future models. To characterize the interactions between group representations and ML models, we construct a participation dynamics model in Chapter 4,

where individuals respond to perceived decisions by leaving ML system uniformly at random: individuals who perceive mistreatment from the decisions are more likely to leave. We aim to understand how group representation disparity evolves in a sequential framework and how enforcing fairness constraints plays a role in this process [153].

2. *Qualification Dynamics* (Chapter 5)
   ML decisions made in the past may affect individuals' true labels in the future ML systems. Consider an example in recruitment where a company aims to select individuals from applicants that are qualified for job positions. Individuals after receiving hiring decisions may take certain actions (e.g., exerting efforts, imitating others, etc.), which results in changes in their future qualifications/labels. As such, the qualification rate–the fraction of the qualified individuals–of each group changes accordingly. To characterize the interactions between qualification rates and ML models, we construct a qualification dynamics model in Chapter 5, where individuals respond to perceived decisions by changing their future qualifications. We aim to understand how group qualification disparity evolves in a sequential framework and how it is affected when various fairness constraints are imposed [155].
3. *Strategic Interplay* (Chapter 6)
   When ML algorithms are used to make high-stake decisions about people, the need for transparency increases in terms of how decisions are reached given input. Given (partial) information about the ML algorithm, individuals may adapt their behavior by strategically manipulating their data to receive favorable decisions. For instance, a hiring or admissions practice that heavily depends on GPA might motivate students to cheat on exams; not accounting for such manipulation may result in disproportionate hiring of under-qualified individuals. A strategic decision maker who anticipates such behavior aims to make its ML models robust to such strategic manipulation. In Chapter 6, we adopt a typical two-stage (Stackelberg) game setting to characterize the interactions between ML models and strategic individuals, where the decision maker commits to its policies, following which individuals best-respond. We aim to examine the impact of the anticipation of strategic behavior and understand whether fairness interventions can serve as incentives/disincentives for strategic manipulation.

Chapter 4: Long-Term Impact of (Fair) ML on Group Representation. Machine Learning (ML) models trained on data from multiple demographic groups can inherit representation disparity

that may exist in the data: the model may be less favorable to groups contributing less to the training process; this in turn can degrade population retention in these groups over time, and exacerbate representation disparity in the long run. In this chapter, we seek to understand the interplay between ML decisions and the underlying group representation, how they evolve in a sequential framework, and how the use of fairness criteria plays a role in this process. To this end, we first construct a participation dynamics model to characterize the interplay between ML decisions and group representations, where individual's retention/participation is driven by the mistreatment perceived from the decision, i.e., individuals who perceive mistreatment from the decisions are more likely to leave the system. Under such dynamics, we conduct equilibrium analysis and study the long-term impact of fairness interventions by comparing the equilibria under different (fair) ML decisions. Our results show that the representation disparity can easily worsen over time when decisions are made based on a commonly used objective and fairness criteria, resulting in some groups diminishing entirely from the sample pool in the long run. It highlights the fact that fairness criteria have to be defined while taking into consideration the impact of decisions on user dynamics. Furthermore, we introduce an approach to selecting a proper fairness criterion based on a general dynamics model, which can balance the group representations in the long run.

Chapter 5: Long-Term Impact of (Fair) ML on Group Qualification. In this chapter, we examine the interplay between ML models and underlying group qualifications, and we aim to understand how the group qualification disparity evolves in a sequential framework and how it is affected under various fairness interventions. To this end, we use a Partially Observed Markov Decision Process (POMDP) framework to formulate the qualification dynamics model, where the unqualified/qualified individuals in the past can become qualified/unqualified with some decision-dependent probabilities. The model indicates how individuals' qualifications transition over two consecutive time steps and characterizes the interplay between ML decisions and group qualifications. Under such dynamics, we conduct the equilibrium analysis and identify conditions for the existence of an unique equilibrium. Furthermore, we examine the long-term impact of fairness interventions on the group qualification disparity by comparing the equilibria under different (fair) ML decisions. Our results show that fairness interventions can either promote equality or exacerbate disparity depending on the qualification transitions and the effect of group sensitive attributes on feature distributions. We also consider possible interventions that can effectively improve group qualification or promote equality of group qualification. Our theoretical results and experiments on

static real-world datasets with simulated dynamics show that our framework can be used to facilitate social science studies.

Chapter 6: Impact of (Fair) ML on Strategic Manipulation. In this chapter, we study fairness issues in the presence of individual's strategic behavior. We suppose individuals can observe ML models used by a decision maker in advance and they can manipulate their data strategically to receive favorable decisions. We aim to design (fair) algorithms that are robust to strategic manipulation, and to understand the impact of fairness interventions on individual's strategic manipulative behavior. To this end, we use a two-stage (Stackelberg) game to characterize the interactions between decision makers and individuals, where the the former first publishes the ML models and the latter may manipulate their features in order to receive more favorable decisions. Depending on whether the decision-maker can anticipate such strategic manipulation or not, the models can be strategic or non-strategic. Moreover, the models may or may not satisfy a fairness constraint. We analytically characterize the equilibrium strategies of both decision maker and individuals, and examine how the algorithms and their resulting fairness properties are affected when the decision maker is strategic (anticipates manipulation), as well as the impact of fairness interventions on equilibrium strategies. In particular, we identify conditions under which anticipation of strategic behavior may mitigate/exacerbate unfairness, and conditions under which fairness interventions can serve as incentives/disincentives for strategic manipulation.

Part I

Designing Differentially Private Algorithms

CHAPTER 2

Private ADMM-Based Distributed Algorithms

2.1 Introduction

In this chapter, we design differentiially private algorithms for distributed optimization. Distributed optimization and learning are crucial for many settings where where the data is possessed by multiple parties or when the quantity of data prohibits processing at a central location. It helps to reduce the computational complexity, improve both the robustness and the scalability of data processing. Many problems can be formulated as a convex optimization of the following form:

$$\min_{\mathbf{x}} \sum_{i=1}^{N} f_i(\mathbf{x}).$$

In a distributed setting, each entity/node i has its own local objective f_i, N entities/nodes collaboratively work to solve this objective through an interactive process of local computation and message passing, which ideally should result in all nodes converging to a global optimum.

Existing approaches to decentralizing the above problem primarily consist of subgradient-based algorithms [48,106,116], ADMM-based algorithms [100,101,124,134,139,140,146], and composite of subgradient and ADMM [14]. It has been shown that ADMM-based algorithms can converge at the rate of $O(\frac{1}{k})$ while subgradient-based algorithms typically converge at the rate of $O(\frac{1}{\sqrt{k}})$, where k is the number of iterations [134]. In this study, we will solely focus on ADMM-based algorithms.

The information exchanged over the iterative process gives rise to privacy concerns if the local training data is proprietary to each node, especially when it contains sensitive information such as medical or financial records, web search history, and so on [89, 130]. It is therefore highly desirable to ensure such iterative processes are privacy-preserving.

A widely used notion of privacy is the ε-differential privacy; it is generally achieved by perturbing the algorithm such that the probability distribution of its output is relatively insensitive to any change to a single record in the input [35]. Several differentially private distributed algorithms have been proposed, including [13, 54, 55, 68, 147]. While a number of such studies have been done for (sub)gradient-based algorithms [13, 54, 55, 68], the same is much harder for ADMM-based algorithms due to its computational complexity stemming from the fact that each node is required to solve an optimization problem in each iteration. To the best of our knowledge, only [147] applies differential privacy to ADMM, where the noise is either added to the dual variable (*dual variable perturbation*) or the primal variable (*primal variable perturbation*) in ADMM updates. However, [147] could only bound the privacy loss of a single iteration. Since an attacker can potentially use all intermediate results to perform inference, the privacy loss accumulates over time through the iterative process. It turns out that the tradeoff between the accuracy of the algorithm and its privacy preservation over the entire computational process becomes hard using the existing method.

In this chapter, we address those issues by inspecting the total privacy loss over the entire process and the entire network. We further propose a number of privacy-preserving algorithms that could simultaneously improve the accuracy and privacy for ADMM. In particular, We have explored two ideas when designing algorithms:

(a) Improve algorithmic robustness to accommodate more randomness.

(b) Reuse intermediate computational results to reduce the total information leakage.

Our main contributions are as follows.

1. We employ idea (a) and propose modified ADMM (M-ADMM) whereby each node independently decides its own penalty parameter in each iteration; it may also differ from the dual updating step size (Section 2.3.1).
2. We employ idea (b) and propose recycled ADMM (R-ADMM) whereby the computational outcomes during even updates are repeatedly used for odd updates; it ensures that the odd updates incur no privacy loss and require much less computation (Section 2.3.2).
3. We employ both ideas and propose modified recycled ADMM (MR-ADMM), where we generalize R-ADMM by accommodating time-varied, node-dependent penalty parameters (Section 2.3.3).

4. We establish sufficient conditions for convergence of M-ADMM, R-ADMM, and MR-ADMM, and quantify the lower bound of the convergence rate for M-ADMM (Section 2.4).
5. We present perturbation mechanisms to provide differential privacy for M-ADMM, R-ADMM, and MR-ADMM (Section 2.5), and characterize the total privacy loss for these private algorithms (Section 2.6). We quantify the generalization performance of (private) MR-ADMM by conducting sample complexity analysis (Section 2.7).
6. We conduct experiments on real-world data (Section 2.9), the empirical results show that our proposed algorithms can achieve stronger privacy guarantee as well as better algorithmic performance, i.e., more stable convergence and higher accuracy.

The remainder of the chapter is organized as follows. We present problem formulation and the definition of differential privacy and ADMM in Section 2.2. Three algorithms are introduced in Section 2.3 including M-ADMM, R-ADMM and MR-ADMM. The convergence analysis of three algorithms are presented in Section 2.4. The private versions of these algorithms, privacy analysis, and sample complexity analysis are presented in Sections 2.5, 2.6, and 2.7, respectively. Discussion is given in Section 2.8. Numerical results are illustrated in Section 2.9. All proofs can be found in Appendix A.

2.2 Preliminaries

2.2.1 Problem Formulation

Consider a connected network[1] given by an undirected graph $G(\mathcal{N}, \mathcal{E})$, which consists of a set of nodes $\mathcal{N} = \{1, 2, \cdots, N\}$ and a set of edges $\mathcal{E} = \{1, 2, \cdots, E\}$. Two nodes can exchange information if and only if they are connected by an edge. Let $\mathcal{V}_i$ denote node i's set of neighbors, excluding itself. Let D_i be node i's dataset.

Consider an optimization problem over this network of N nodes, where the overall objective function can be decomposed into N sub-objective functions and each depends on a node's local dataset, i.e.,

$$\min_{f_c} \text{Obj}(f_c, D_{all}) = \sum_{i=1}^{N} O(f_c, D_i) \tag{2.1}$$

[1] A connected network is one in which every node is reachable (via a path) from every other node.

The goal is to find a (centralized) optimal solution $f_c \in \mathbb{R}^d$ over the union of all local datasets $D_{all} = \cup_{i \in \mathcal{N}} D_i$ in a distributed manner using ADMM, while providing privacy guarantee for each data sample.

2.2.2 Differential Privacy in Optimization

The definition of (ϵ, δ)-differential privacy is formally introduced in Chapter 1.1.1. In this chapter, we adopt pure ϵ-differential privacy when $\delta = 0$, although a weaker notion (ϵ, δ)-differential privacy can also be adopted. This is discussed in Section 2.8.

For an optimization problem over a dataset, there are many approaches to randomizing the output to preserve differential privacy and some of the most commonly used are as follows.

1. *Output perturbation:* solve the optimization problem first and then add zero-mean noise (e.g., Laplace, Gaussian) to the optimal solution.
2. *Objective perturbation:* add a noisy term to the objective function first and then solve the perturbed optimization problem.

Because of this randomness, the accuracy of the output also decreases accordingly. The more perturbation, the output will be less accurate but it also provides the stronger privacy for individuals. Therefore, there is a privacy-accuracy trade-off, and an important issue is how to improve this trade-off so that the output can be more accurate under the same privacy guarantee.

2.2.3 Conventional ADMM

To decentralize (2.1), let f_i be the local classifier of each node i. To achieve consensus, i.e., $f_1 = f_2 = \cdots = f_N$, a set of auxiliary variables $\{w_{ij} | i \in \mathcal{N}, j \in \mathcal{V}_i\}$ are introduced for every pair of connected nodes. As a result, (2.1) is reformulated equivalently as:

$$\begin{aligned} \min_{\{f_i\},\{w_{ij}\}} \quad & \widetilde{\text{Obj}}(\{f_i\}_{i=1}^N, D_{all}) = \sum_{i=1}^{N} O(f_i, D_i) \\ \text{s.t.} \quad & f_i = w_{ij}, w_{ij} = f_j, \quad i \in \mathcal{N}, j \in \mathcal{V}_i \end{aligned} \tag{2.2}$$

Let $\{f_i\}$ and $\{w_{ij}\}$ be the shorthand for $\{f_i\}_{i \in \mathcal{N}}$ and $\{w_{ij}\}_{i \in \mathcal{N}, j \in \mathcal{V}_i}$, respectively. Let $\{w_{ij}, \lambda_{ij}^k\}$ be the shorthand for $\{w_{ij}, \lambda_{ij}^k\}_{i \in \mathcal{N}, j \in \mathcal{V}_i, k \in \{a,b\}}$, where λ_{ij}^a, λ_{ij}^b are dual variables corresponding to equality

constraints $f_i = w_{ij}$ and $w_{ij} = f_j$ respectively. The objective in (2.2) can be solved using ADMM with the augmented Lagrangian:

$$\begin{aligned} L_\eta(\{f_i\},\{w_{ij},\lambda_{ij}^k\}) &= \sum_{i=1}^{N} O(f_i,D_i) + \sum_{i=1}^{N}\sum_{j\in\mathscr{V}_i}(\lambda_{ij}^a)^T(f_i - w_{ij}) + \sum_{i=1}^{N}\sum_{j\in\mathscr{V}_i}(\lambda_{ij}^b)^T(w_{ij}-f_j) \\ &\quad + \sum_{i=1}^{N}\sum_{j\in\mathscr{V}_i}\frac{\eta}{2}(\|f_i - w_{ij}\|_2^2 + \|w_{ij}-f_j\|_2^2)\,. \end{aligned} \tag{2.3}$$

where η is called the penalty parameter. In the $(t+1)$-th iteration, the ADMM updates consist of the following:

$$f_i(t+1) = \underset{f_i}{\operatorname{argmin}}\, L_\eta(\{f_i\},\{w_{ij}(t),\lambda_{ij}^k(t)\}); \tag{2.4}$$

$$w_{ij}(t+1) = \underset{w_{ij}}{\operatorname{argmin}}\, L_\eta(\{f_i(t+1)\},\{w_{ij},\lambda_{ij}^k(t)\}); \tag{2.5}$$

$$\lambda_{ij}^a(t+1) = \lambda_{ij}^a(t) + \eta(f_i(t+1) - w_{ij}(t+1)); \tag{2.6}$$

$$\lambda_{ij}^b(t+1) = \lambda_{ij}^b(t) + \eta(w_{ij}(t+1) - f_j(t+1)). \tag{2.7}$$

Using Lemma 3 in [45], if dual variables $\lambda_{ij}^a(t)$ and $\lambda_{ij}^b(t)$ are initialized to zero for all node pairs (i,j), then $\lambda_{ij}^a(t) = \lambda_{ij}^b(t)$ and $\lambda_{ij}^k(t) = -\lambda_{ji}^k(t)$ will hold for all iterations with $k \in \{a,b\}, i \in \mathscr{N}, j \in \mathscr{V}_i$. Let $\lambda_i(t) = \sum_{j\in\mathscr{V}_i}\lambda_{ij}^a(t) = \sum_{j\in\mathscr{V}_i}\lambda_{ij}^b(t)$, then the ADMM iterations (2.4)-(2.7) can be simplified as:

$$f_i(t+1) = \underset{f_i}{\operatorname{argmin}}\{O(f_i,D_i) + 2\lambda_i(t)^T f_i + \eta\sum_{j\in\mathscr{V}_i}\|\frac{1}{2}(f_i(t)+f_j(t)) - f_i\|_2^2\}\,; \tag{2.8}$$

$$\lambda_i(t+1) = \lambda_i(t) + \frac{\eta}{2}\sum_{j\in\mathscr{V}_i}(f_i(t+1) - f_j(t+1))\,. \tag{2.9}$$

2.2.4 Private ADMM Proposed in [147]

Two randomizations were proposed in [147]:

1. *Dual variable perturbation:* each node i adds a random noise to its dual variable $\lambda_i(t)$ before updating its primal variable $f_i(t)$ using (2.8) in each iteration.
2. *Primal variable perturbation:* after updating primal variable $f_i(t)$, each node adds a random noise to it before broadcasting to its neighbors.

In both methods, the privacy property is only evaluated for a single node and a single iteration. Neither can effectively balance the privacy-accuracy tradeoff if the total privacy loss is considered. In contrast, we consider the total privacy loss of the whole network over the entire iterative process and propose multiple algorithms under which the trade-off between total privacy loss and accuracy can be improved significantly.

2.3 Proposed Algorithms

In this section, we introduce three variants of ADMM algorithm.

2.3.1 Modified ADMM (M-ADMM): Making η a Node's Private Information

Conventional ADMM [18] requires that the penalty parameter η be fixed and equal to the dual updating step size for all nodes in all iterations. Varying the penalty parameter to accelerate convergence in ADMM has been proposed in the literature. For instance, [9, 61, 108, 139] vary this penalty parameter in every iteration but keep it the same for different equality constraints in (2.2). In [125, 145] this parameter varies in each iteration and is allowed to differ for different equality constraints. However, all of these modifications are based on the original ADMM (Eqn. (2.4)-(2.7)) and not on the simplified version (Eqn. (2.8)-(2.9)); the significance of this difference is discussed below in the context of privacy requirement. Moreover, we will decouple $\eta_i(t+1)$ from the dual updating step size, denoted as θ below. For simplicity, θ is fixed for all nodes in our analysis, but can also be private information as we show in numerical experiments.

First consider replacing η with $\eta_{ij}(t+1)$ in Eqn. (2.4)-(2.5) of the original ADMM (as is done in [125, 145]) and replacing η with θ in Eqn. (2.6)-(2.7); we obtain the following:

$$\begin{aligned} f_i(t+1) &= \underset{f_i}{\operatorname{argmin}} \{O(f_i, D_i) + 2\lambda_i(t)^T f_i + \sum_{j\in\mathscr{V}_i} \frac{\eta_{ij}(t+1)+\eta_{ji}(t+1)}{2} \|\frac{1}{2}(f_i(t)+f_j(t)) - f_i\|_2^2\} ; \\ \lambda_i(t+1) &= \lambda_i(t) + \frac{\theta}{2}\sum_{j\in\mathscr{V}_i}(f_i(t+1) - f_j(t+1)) . \end{aligned}$$

This however violates our requirement that $\eta_{ji}(t)$ be node j's private information since this is needed by node i to perform the above computation. To resolve this, we instead start from the simplified

ADMM, modifying Eqn. (2.8)-(2.9):

$$f_i(t+1) = \underset{f_i}{\operatorname{argmin}} \{O(f_i, D_i) + 2\lambda_i(t)^T f_i + \eta_i(t+1) \sum_{j \in \mathcal{V}_i} \|f_i - \frac{1}{2}(f_i(t) + f_j(t))\|_2^2 \} ; \quad (2.10)$$

$$\lambda_i(t+1) = \lambda_i(t) + \frac{\theta}{2} \sum_{j \in \mathcal{V}_i} (f_i(t+1) - f_j(t+1)) , \quad (2.11)$$

where $\eta_i(t+1)$ is now node i's private information. Indeed $\eta_i(t+1)$ is no longer purely a penalty parameter related to any equality constraint in the original sense. We will however refer to it as the private penalty parameter for simplicity. The above constitutes the M-ADMM algorithm.

The penalty parameter $\eta_i(t+1)$ directly controls the step size of the algorithm. Since the goal is to minimize the objective in (2.10), if $\eta_i(t+1)$ is larger, the solution $f_i(t+1)$ will be closer to the primal variable in the previous iteration so that the penalty term $\sum_{j \in \mathcal{V}_i} \|\frac{1}{2}(f_i(t) + f_j(t)) - f_i\|_2^2$ will be small. In other words, larger $\eta_i(t+1)$ results in smaller update of the primal variable $f_i(t+1)$. Therefore, increasing $\eta_i(t+1)$ decreases the step sizes.

Without perturbation, decreasing step size might slow down the convergence. However, when the algorithm is perturbed with added noise, a smaller step size could prevent the variable from deviating too much from the optimal solution in each update, which in turn stabilizes the algorithm.

2.3.2 Recycled ADMM (R-ADMM): Making Information Recyclable

ADMM can outperform gradient-based methods in terms of requiring fewer number of iterations for convergence; this however comes at the price of high computational cost in every iteration. In particular, the primal variable is updated by performing an optimization in each iteration. In [101, 102, 112], either a linear or quadratic approximation of the objective function is used to obtain an inexact solution in each iteration in lieu of solving the original optimization problem. While this clearly lowers the computational cost, the approximate computation is performed using the local, individual data in every iteration, which means that privacy loss inevitably accumulates over the iterations.

We modify ADMM in such a way that in every even iteration, without using individual's data D_{all}, the primal variable is updated solely based on the existing computational results from the previous, odd iteration. Compared with conventional ADMM, these updates incur no privacy loss and less computation. Since the computational results are repeatedly used, this method is referred

to as Recycled ADMM (R-ADMM).

Specifically, in the $2k$-th (even) iteration, $O(f_i, D_i)$ (Eqn. (2.8), primal update optimization) is approximated by

$$O(f_i, D_i) \approx O(f_i(2k-1), D_i) + \nabla O(f_i(2k-1), D_i)^T (f_i - f_i(2k-1)) + \frac{\gamma}{2}\|f_i - f_i(2k-1)\|_2^2$$

where $\gamma \geq 0$. Moreover, only the primal variables are updated in the $2k$-th (even) iteration. Using the first-order condition, the updates in the $2k$-th iteration become:

$$\begin{aligned} f_i(2k) &= f_i(2k-1) - \frac{1}{2\eta V_i + \gamma}\{\nabla O(f_i(2k-1), D_i) + 2\lambda_i(2k-1) \\ &\quad + \eta \sum_{j \in \mathcal{V}_i} (f_i(2k-1) - f_j(2k-1))\}\ ; \qquad (2.12) \\ \lambda_i(2k) &= \lambda_i(2k-1)\ . \qquad (2.13) \end{aligned}$$

In the $(2k-1)$-th (odd) iteration, the updates are kept the same as (2.8)(2.9):

$$\begin{aligned} f_i(2k-1) &= \underset{f_i}{\operatorname{argmin}}\{O(f_i, D_i) + 2\lambda_i(2k-2)^T f_i \\ &\quad + \eta \sum_{j \in \mathcal{V}_i} \|\frac{1}{2}(f_i(2k-2) + f_j(2k-2)) - f_i\|_2^2\ \}\ ; \qquad (2.14) \\ \lambda_i(2k-1) &= \lambda_i(2k-2) + \frac{\eta}{2} \sum_{j \in \mathcal{V}_i} (f_i(2k-1) - f_j(2k-1))\ . \qquad (2.15) \end{aligned}$$

Note that in the $(2k)$-th (even) iteration, we need the gradient $\nabla O(f_i(2k-1), D_i)$ and primal difference $\frac{\eta}{2}\sum_{j \in \mathcal{V}_i}(f_i(2k-1) - f_j(2k-1))$ for the updates; these are available directly from the previous, $(2k-1)$-th (odd) iteration, i.e., this information can be recycled. In this sense, R-ADMM may be viewed as alternating between conventional ADMM (odd iterations) and a variant of gradient descent (even iterations), where $\frac{1}{2\eta V_i + \gamma}$ is the step-size with a slightly modified gradient term.

2.3.3 Modified R-ADMM (MR-ADMM): M-ADMM + R-ADMM

R-ADMM requires that the penalty parameter η be fixed for all nodes in all iterations. We can further implement idea in M-ADMM by modifying R-ADMM such that each node can independently determine its penalty parameter in each iteration. Specifically, replace η in (2.12), (2.14) and (2.15)

with $\eta_i(2k-1)$. The updating formula is then given in (2.16)-(2.19).

$$
\begin{aligned}
f_i(2k-1) &= \underset{f_i}{\operatorname{argmin}}\{O(f_i, D_i) + 2\lambda_i(2k-2)^T f_i \\
&\quad + \eta_i(2k-1)\sum_{j\in\mathscr{V}_i} \|\frac{1}{2}(f_i(2k-2) + f_j(2k-2)) - f_i\|_2^2 \}\ ; \qquad (2.16)\\
\lambda_i(2k-1) &= \lambda_i(2k-2) + \frac{\eta_i(2k-1)}{2}\sum_{j\in\mathscr{V}_i}(f_i(2k-1) - f_j(2k-1))\ . \qquad (2.17)\\
f_i(2k) &= f_i(2k-1) - \frac{1}{2\eta_i(2k-1)V_i + \gamma}\{\nabla O(f_i(2k-1), D_i) + 2\lambda_i(2k-1) \\
&\quad + \eta_i(2k-1)\sum_{j\in\mathscr{V}_i}(f_i(2k-1) - f_j(2k-1))\}\ ; \qquad (2.18)\\
\lambda_i(2k) &= \lambda_i(2k-1)\ . \qquad (2.19)
\end{aligned}
$$

Note that MR-ADMM is a generalized version of R-ADMM. If fix $\eta_i(2k-1) = \eta$, $\forall k$, then MR-ADMM will be reduced to R-ADMM.

2.4 Convergence Analysis

In this section, we show that M-ADMM (Eqn. (2.10)-(2.11)), R-ADMM (Eqn. (2.12)-(2.15)), and MR-ADMM (Eqn. (2.16)-(2.19)) all converge to the optimal solution under a set of common technical assumptions.

Assumption 1. *Function $O(f_i, D_i)$ is convex and continuously differentiable in f_i, $\forall i$.*

Assumption 2. *The solution set to the original ERM problem* (2.1) *is nonempty and there exists at least one bounded element.*

Define the adjacency matrix of the network $A \in \mathbb{R}^{N\times N}$ as

$$
a_{ij} = \begin{cases} 1, & \text{if node } i \text{ and node } j \text{ are connected} \\ 0, & \text{otherwise}\ . \end{cases}
$$

Stack the variables $f_i(t)$, $\lambda_i(t)$ and $\nabla O(f_i(t), D_i)$ for $i \in \mathcal{N}$ into matrices, i.e.,

$$\hat{f}(t) = \begin{bmatrix} f_1(t)^T \\ f_2(t)^T \\ \vdots \\ f_N(t)^T \end{bmatrix} \in \mathbb{R}^{N\times d},\ \Lambda(t) = \begin{bmatrix} \lambda_1(t)^T \\ \lambda_2(t)^T \\ \vdots \\ \lambda_N(t)^T \end{bmatrix} \in \mathbb{R}^{N\times d},\ \nabla\hat{O}(\hat{f}(t), D_{all}) = \begin{bmatrix} \nabla O(f_1(t), D_1)^T \\ \nabla O(f_2(t), D_2)^T \\ \vdots \\ \nabla O(f_N(t), D_N)^T \end{bmatrix} \in \mathbb{R}^{N\times d}$$

Let $V_i = |\mathcal{V}_i|$ be the number of neighbors of node i, and define the degree matrix $D = \mathbf{diag}([V_1; V_2; \cdots; V_N]) \in \mathbb{R}^{N\times N}$. Note that $D - A$ is the Laplacian matrix and $D + A$ is the signless Laplacian matrix of the network, with the following properties if the network is connected: (i) $D \pm A \succeq 0$ is positive semi-definite; (ii) $\text{Null}(D - A) = c\mathbf{1}$, i.e., every member in the null space of $D - A$ is a scalar multiple of $\mathbf{1}$ with $\mathbf{1}$ being the vector of all 1's [82].

Lemma 1. *[First-order Optimality Condition [100]] Under Assumptions 1 and 2, the following three statements are equivalent:*

- $\hat{f}^* = [(f_1^*)^T; (f_2^*)^T; \cdots; (f_N^*)^T] \in \mathbb{R}^{N\times d}$ *is consensual, i.e.,* $f_1^* = f_2^* = \cdots = f_N^* = f_c^*$ *where* f_c^* *is the optimal solution to* (2.1)*.*
- *There exists a pair* $(\hat{f}^*, Y^*)$ *with* $Y^* = \sqrt{D - A}X$ *for some* $X \in \mathbb{R}^{N\times d}$ *such that*

$$\nabla\hat{O}(\hat{f}^*, D_{all}) + \sqrt{D - A}Y^* = \mathbf{0}_{N\times d}\ ; \tag{2.20}$$

$$\sqrt{D - A}\hat{f}^* = \mathbf{0}_{N\times d}\ . \tag{2.21}$$

- *There exists a pair* $(\hat{f}^*, \Lambda^*)$ *with* $2\Lambda^* = (D - A)X$ *for some* $X \in \mathbb{R}^{N\times d}$ *such that*

$$\nabla\hat{O}(\hat{f}^*, D_{all}) + 2\Lambda^* = \mathbf{0}_{N\times d}\ ; \tag{2.22}$$

$$(D - A)\hat{f}^* = \mathbf{0}_{N\times d}\ . \tag{2.23}$$

2.4.1 M-ADMM

The KKT optimality condition of the primal update (2.10) is:

$$0 = \nabla O(f_i(t+1), D_i) + 2\lambda_i(t) + \eta_i(t+1)\sum_{j\in\mathcal{V}_i}(2f_i(t+1) - (f_i(t) + f_j(t)))\ . \tag{2.24}$$

Define penalty-weighted matrix $W(t) = \mathbf{diag}([\eta_1(t); \eta_2(t); \cdots; \eta_N(t)]) \in \mathbb{R}^{N\times N}$ for t-th iteration. Then the matrix form of (2.11), (2.24) are:

$$\begin{aligned} \mathbf{0}_{N\times d} &= \nabla \hat{O}(\hat{f}(t+1), D_{all}) + 2\Lambda(t) + 2W(t+1)D\hat{f}(t+1) - W(t+1)(D+A)\hat{f}(t) \quad (2.25) \\ 2\Lambda(t+1) &= 2\Lambda(t) + \theta(D-A)\hat{f}(t+1) \quad (2.26) \end{aligned}$$

Let $\sqrt{X}$ denote the square root of a symmetric positive semi-definite (PSD) matrix X that is also symmetric PSD, i.e., $\sqrt{X}\sqrt{X} = X$. Define matrix $Y(t)$ such that $2\Lambda(t) = \sqrt{D-A}Y(t)$. Since $\Lambda(0) = \mathbf{zeros}(N, d)$, which is in the column space of $D-A$, this together with (2.26) imply that $\Lambda(t)$ is in the column space of $D-A$ and $\sqrt{D-A}$. This guarantees the existence of $Y(t)$. This allows us to rewrite (2.25)-(2.26) as:

$$\begin{aligned} \mathbf{0}_{N\times d} &= \nabla \hat{O}(\hat{f}(t+1), D_{all}) + \sqrt{D-A}Y(t+1) + (W(t+1) - \theta I)(D-A)\hat{f}(t+1) \\ &\quad + W(t+1)(D+A)(\hat{f}(t+1) - \hat{f}(t)) \; ; \quad (2.27) \\ Y(t+1) &= Y(t) + \theta\sqrt{D-A}\hat{f}(t+1) \; . \quad (2.28) \end{aligned}$$

Lemma 1 shows that a pair $(Y^*, \hat{f}^*)$ satisfying (2.20)(2.21) is equivalent to the optimal solution of our problem, hence the convergence of M-ADMM is proved by showing that $(Y(t), \hat{f}(t))$ converges to a pair $(Y^*, \hat{f}^*)$ satisfying (2.20)(2.21).

Theorem 1. *Consider the modified ADMM defined by* (2.10)-(2.11). *Let* $\{Y(t), \hat{f}(t)\}$ *be outputs in each iteration and* $(Y^*, \hat{f}^*)$ *a pair satisfying* (2.20)-(2.21). *Denote*

$$Z(t) = \begin{bmatrix} Y(t) \\ \hat{f}(t) \end{bmatrix} \in \mathbb{R}^{2N\times d}, \; Z^* = \begin{bmatrix} Y^* \\ \hat{f}^* \end{bmatrix} \in \mathbb{R}^{2N\times d}, \; J(t) = \begin{bmatrix} \frac{I_{N\times N}}{\theta} & 0 \\ 0 & W(t)(D+A) \end{bmatrix} \in \mathbb{R}^{2N\times 2N}$$

Let $\langle\cdot,\cdot\rangle_F$ *be the Frobenius inner product of two matrices. We have*

$$\langle Z(t+1) - Z^*, J(t+1)(Z(t+1) - Z(t))\rangle_F \leq 0 \; . \quad (2.29)$$

If $\eta_i(t+1) \geq \eta_i(t) \geq \theta > 0$ *and* $\eta_i(t) < +\infty$, $\forall t, i$, *then* $(Y(t), \hat{f}(t))$ *converges to* $(Y^*, \hat{f}^*)$.

Convergence Rate Analysis. To further establish the convergence rate of modified ADMM, an additional assumption is used:

Assumption 3. *For all $i \in \mathcal{N}$, $O(f_i, D_i)$ is strongly convex in f_i and has Lipschitz continues gradients, i.e., for any f_i^1 and f_i^2, we have:*

$$(f_i^1 - f_i^2)^T(\nabla O(f_i^1, D_i) - \nabla O(f_i^2, D_i)) \geq m_i \|f_i^1 - f_i^2\|_2^2$$

$$\|\nabla O(f_i^1, D_i) - \nabla O(f_i^2, D_i)\|_2 \leq M_i \|f_i^1 - f_i^2\|_2 \tag{2.30}$$

where $m_i > 0$ is the strong convexity constant and $0 < M_i < +\infty$ is the Lipschitz constant.

Theorem 2. *Define $D_m = \boldsymbol{diag}([m_1; m_2; \cdots; m_N]) \in \mathbb{R}^{N\times N}$ and $D_M = \boldsymbol{diag}([M_1^2; M_2^2; \cdots; M_N^2]) \in \mathbb{R}^{N\times N}$ with $m_i > 0$ and $0 < M_i < +\infty$ as given in Assumption 3. Denote by $\|X\|_J^2 = \langle X, JX\rangle_F$ the Frobenius inner product of any matrix X and JX; denote by $\sigma_{min}(\cdot)$ and $\sigma_{max}(\cdot)$ the smallest nonzero, and the largest, singular values of a matrix, respectively.*

Let $\tilde{\sigma}_{max}(t) = \sigma_{max}(W(t)(D+A))$, $\bar{\sigma}_{max/min}(t) = \sigma_{max/min}((W(t)-\theta I)(D-A))$ and $\mu > 1$ be an arbitrary constant. Consider any $\delta(t)$ that satisfies (2.31)(2.32)*:*

$$\frac{\delta(t)\mu^2\tilde{\sigma}_{max}(t)}{\theta\sigma_{min}(D-A)} \leq 1 \tag{2.31}$$

and

$$\delta(t)\left(\frac{\mu\bar{\sigma}_{max}(t)^2\boldsymbol{I}_N + \mu^2 D_M}{\theta\sigma_{min}(D-A)(\mu-1)} + W(t)(D+A)\right) \leq 2(W(t)-\theta I)(D-A) + 2D_m\ . \tag{2.32}$$

If $\eta_i(t+1) \geq \eta_i(t) \geq \theta > 0$ and $\eta_i(t) < +\infty$, $\forall t, i$, then $(Y(t), \hat{f}(t))$ converges to $(Y^*, \hat{f}^*)$ in the following sense:

$$(1+\delta(t))\|Z(t) - Z^*\|_{J(t)}^2 \leq \|Z(t-1) - Z^*\|_{J(t)}^2\ .$$

Furthermore, a lower bound on $\delta(t)$ is:

$$\min\{\frac{\theta\sigma_{min}(D-A)}{\mu^2\tilde{\sigma}_{max}(t)}, \frac{2m_o + 2\bar{\sigma}_{min}(t)}{\frac{\mu^2 M_O^2 + \mu\bar{\sigma}_{max}(t)^2}{\theta\sigma_{min}(D-A)(\mu-1)} + \tilde{\sigma}_{max}(t)}\} \tag{2.33}$$

where $m_o = \min_{i\in\mathcal{N}}\{m_i\}$ and $M_O = \max_{i\in\mathcal{N}}\{M_i\}$.

Although Theorem 2 only gives a lower bound on the convergence rate $(1+\delta(t))$ of the M-ADMM, it reflects the impact of penalty $\{\eta_i(t)\}_{i=1}^N$ on the convergence. Since $\bar{\sigma}_{\max}(t) = \sigma_{\max}((W(t)-\theta I)(D-A))$ and $\tilde{\sigma}_{\max}(t) = \sigma_{\max}(W(t)(D+A))$, larger penalty results in larger $\bar{\sigma}_{\max}(t)$ and $\tilde{\sigma}_{\max}(t)$. By (2.33), the first term, $\frac{\theta\sigma_{\min}(D-A)}{\mu^2\tilde{\sigma}_{\max}(t)}$ is smaller when $\tilde{\sigma}_{\max}(t)$ is larger. The second term

is bounded by $\frac{\theta\sigma_{\min}(D-A)(\mu-1)(2m_o+2\bar{\sigma}_{\min}(t))}{\mu\bar{\sigma}_{\max}(t)^2}$, which is smaller when $\bar{\sigma}_{\max}(t)$ is larger. Therefore, the convergence rate $1+\delta(t)$ decreases as $\{\eta_i(t)\}_{i=1}^N$ increase.

2.4.2 R-ADMM & MR-ADMM

Since MR-ADMM is a generalized version of R-ADMM, we focus on the convergence analysis of MR-ADMM in this section while the results immediately apply to R-ADMM by fixing $\eta_i(2k-1)=\eta$, $\forall k$. To prove the convergence of MR-ADMM (Eqn. (2.16)-(2.19)), we introduce an additional assumption below.

Assumption 4. *For all $i\in\mathcal{N}$, $O(f_i,D_i)$ has Lipschitz continuous gradients, i.e., for any f_i^1 and f_i^2, we have:*

$$\|\nabla O(f_i^1,D_i)-\nabla O(f_i^2,D_i)\|_2\le M_i\|f_i^1-f_i^2\|_2 \tag{2.34}$$

The KKT condition of the primal update (2.16) is given as:

$$0=\nabla O(f_i(2k-1),D_i)+2\lambda_i(2k-2)+\eta_i(2k-1)\sum_{j\in\mathcal{V}_i}(2f_i(2k-1)-(f_i(2k-2)+f_j(2k-2))). \tag{2.35}$$

Define the diagonal matrix $\tilde{D}(2k-1)$ with $\tilde{D}(2k-1)_{ii}=2\eta_i(2k-1)V_i+\gamma$, and the weight matrix $W(2k-1)=\mathbf{diag}([\eta_1(2k-1);\eta_2(2k-1);\cdots;\eta_N(2k-1)])\in\mathbb{R}^{N\times N}$. Then for each k, the matrix form of (2.18)(2.19)(2.35)(2.17) are given in (2.36)-(2.39):

$$\begin{aligned}
\hat{f}(2k) &= \hat{f}(2k-1)-\tilde{D}(2k-1)^{-1}\{\nabla\hat{O}(\hat{f}(2k-1),D_{all})+2\Lambda(2k-1)\\
&\quad +W(2k-1)(D-A)\hat{f}(2k-1)\}\ ; &(2.36)\\
2\Lambda(2k) &= 2\Lambda(2k-1)\ ; &(2.37)\\
\mathbf{0}_{N\times d} &= \nabla\hat{O}(\hat{f}(2k-1),D_{all})+2\Lambda(2k-2)\\
&\quad +W(2k-1)(2D\hat{f}(2k-1)-(D+A)\hat{f}(2k-2))\ ; &(2.38)\\
2\Lambda(2k-1) &= 2\Lambda(2k-2)+W(2k-1)(D-A)\hat{f}(2k-1)\ . &(2.39)
\end{aligned}$$

Writing $\hat{f}(2k-2)$ and $\Lambda(2k-2)$ in (2.38)(2.39) as functions of $\hat{f}(2k-3)$, $\Lambda(2k-3)$ using

(2.36)(2.37), we obtain Eqn. (2.40)(2.41).

$$\begin{aligned}
\mathbf{0}_{N\times d} &= \nabla\hat{O}(\hat{f}(2k-1), D_{all}) + W(2k-1)(D+A)\tilde{D}(2k-3)^{-1}\nabla\hat{O}(\hat{f}(2k-3), D_{all}) \\
&+ W(2k-1)(D+A)(\hat{f}(2k-1) - \hat{f}(2k-3)) \\
&+ W(2k-1)(D+A)\tilde{D}(2k-3)^{-1}W(2k-3)(D-A)\hat{f}(2k-3) \\
&+ 2\Lambda(2k-1) + W(2k-1)(D+A)\tilde{D}(2k-3)^{-1}2\Lambda(2k-3)\ ; \qquad (2.40) \\
2\Lambda(2k-1) &= 2\Lambda(2k-3) + W(2k-1)(D-A)\hat{f}(2k-1)\ . \qquad (2.41)
\end{aligned}$$

The convergence of MR-ADMM is proved by showing that the pair $(\hat{f}(2k-1), \Lambda(2k-1))$ from odd iterations converges to the optimal solution. To simplify the notation, we will re-index every two consecutive odd iterations $2k-3$ and $2k-1$ using t and $t+1$, it results in Eqn. (2.42)(2.43).

$$\begin{aligned}
\mathbf{0}_{N\times d} &= \nabla\hat{O}(\hat{f}(t+1), D_{all}) + W(t+1)(D+A)\tilde{D}(t)^{-1}\nabla\hat{O}(\hat{f}(t), D_{all}) \\
&\quad + W(t+1)(D+A)(\hat{f}(t+1) - \hat{f}(t)) + W(t+1)(D+A)\tilde{D}(t)^{-1}2\Lambda(t) \\
&+ W(t+1)(D+A)\tilde{D}(t)^{-1}W(t)(D-A)\hat{f}(t) + 2\Lambda(t+1)\ ; \qquad (2.42) \\
2\Lambda(t+1) &= 2\Lambda(t) + W(t+1)(D-A)\hat{f}(t+1)\ . \qquad (2.43)
\end{aligned}$$

Lemma 1 shows that a pair $(\hat{f}^*, \Lambda^*)$ satisfying (2.22)(2.23) is equivalent to the optimal solution of our problem, hence the convergence of the MR-ADMM is proved by showing that $(\hat{f}(t), \Lambda(t))$ in (2.42)(2.43) converges to a pair $(\hat{f}^*, \Lambda^*)$ satisfying (2.22)(2.23).

Theorem 3. ***[Sufficient Condition]*** *Consider the modified ADMM defined by* (2.42)(2.43)*. Let* $\{\hat{f}(t), \Lambda(t)\}$ *be outputs in each iteration and* $\{\hat{f}^*, \Lambda^*\}$ *a pair satisfying* (2.22)(2.23)*. Denote* $D_M = \boldsymbol{diag}([M_1^2; M_2^2; \cdots; M_N^2]) \in \mathbb{R}^{N\times N}$ *with* $0 < M_i < +\infty$ *as given in Assumption 4. If* $\eta_i(t+1) \geq \eta_i(t) > 0$ *and* $\eta_i(t) < +\infty$ *hold and the following two conditions can also be satisfied for some constants* $L > 0$ *and* $\mu > 1$*:*

$$\begin{aligned}
(i)\ I + W(t+1)(D+A)\tilde{D}(t)^{-1} &\succ \frac{L\mu}{2\sigma_{\min}(\tilde{D}(t))}(W(t+1)(D-A))^{+}D_M\ ; \\
(ii)\qquad W(t+1)(D+A) &\succ W(t+1)(D+A)\tilde{D}(t)^{-1}\Big(W(t)(D-A) \\
&\quad + \frac{2}{L}W(t+1)(D+A)\Big) + \frac{L\mu}{2\sigma_{\min}(\tilde{D}(t))(\mu-1)}D_M\ .
\end{aligned}$$

where $\sigma_{\min}(\tilde{D}(t)) = \min_i\{2\eta_i(t)V_i + \gamma\}$ *is the smallest singular value of* $\tilde{D}(t)$*, then* $(\hat{f}(t), \Lambda(t))$ *converges to* $(\hat{f}^*, \Lambda^*)$.

By controlling γ to be sufficiently large, $\tilde{D}(t)_{ii} = 2\eta_i(t)V_i + \gamma$ will be large and conditions *(i)(ii)* can always be satisfied under some constants $L > 0$ and $\mu > 1$. Note that the conditions *(i)(ii)* are sufficient but not necessary, so in practice convergence may be attained under weaker settings.

For R-ADMM, take $L = 2$ and $\mu = 2$, then condition *(i)(ii)* are reduced to:

$$(iii) \quad I + \eta(D+A)\tilde{D}^{-1} \succ \frac{2}{\eta\sigma_{\min}(\tilde{D})}((D-A))^+ D_M \ ;$$

$$(iv) \quad \eta(D+A) \succ 2\eta(D+A)\tilde{D}^{-1}\eta D + \frac{2}{\sigma_{\min}(\tilde{D})} D_M \ .$$

Again for a sufficiently large $\gamma \geq 0$, *(iii)(iv)* can be easily satisfied.

2.5 Private Algorithms

In this section we present privacy preserving versions of M-ADMM and MR-ADMM. Since MR-ADMM is a generalized version of R-ADMM, the private version of R-ADMM can be built in a similar way.

2.5.1 Private M-ADMM

A random noise $\epsilon_i(t+1)$ with probability density proportional to $\exp\{-\alpha_i(t+1)\|\epsilon_i(t+1)\|_2\}$ is added to penalty term in the objective function of (2.10):

$$L_i^{priv}(t+1) = O(f_i, D_i) + 2\lambda_i(t)^T f_i + \eta_i(t+1) \sum_{j \in \mathscr{V}_i} \|f_i + \epsilon_i(t+1) - \frac{1}{2}(f_i(t) + f_j(t))\|_2^2 \tag{2.44}$$

To generate this noisy vector, choose the norm from the gamma distribution with shape d and scale $\frac{1}{\alpha_i(t+1)}$ and the direction uniformly, where d is the dimension of the feature space. Then node i's local result is obtained by finding the optimal solution to the private objective function:

$$f_i(t+1) = \operatorname*{argmin}_{f_i} L_i^{priv}(t+1), \ \ i \in \mathscr{N} \ . \tag{2.45}$$

It is equivalent to (2.48) below when noise $\eta_i(t+1)V_i\epsilon_i(t+1)$ is added to the dual variable $\lambda_i(t)$:

$$\begin{aligned}\underset{f_i}{\operatorname{argmin}}\ \tilde{L}_i^{priv}(t+1) &= O(f_i, D_i) + 2(\lambda_i(t) + \eta_i(t+1)V_i\epsilon_i(t+1))^T f_i \\ &\quad + \eta_i(t+1)\sum_{j\in\mathscr{V}_i} \|f_i - \frac{1}{2}(f_i(t) + f_j(t))\|_2^2 \,.\end{aligned} \tag{2.46}$$

Further, if $\eta_i(t+1) = \eta = \theta, \forall i, t$, then the above is reduced to the dual variable perturbation in [147][2].

The complete procedure is shown in Algorithm 1, where the condition used to generate θ helps bound the worst-case privacy loss but is not necessary in guaranteeing convergence.

Algorithm 1: Private M-ADMM

Input: $\{D_i\}_{i=1}^N$, $\{\alpha_i(1), \cdots, \alpha_i(K)\}_{i=1}^N$
Initialize: Generate $f_i(0)$ randomly and $\lambda_i(0) = \mathbf{0}_{d\times 1}$ for every node $i \in \mathscr{N}$, $t = 0$
Parameter: Determine θ such that $2c_1 < \frac{B_i}{C}(\frac{\rho}{N} + 2\theta V_i)$ holds for all i.
for $t = 1$ **to** T **do**
  for $i = 1$ **to** $\mathscr{N}$ **do**
    Generate noise $\epsilon_i(t+1) \sim \exp(-\alpha_i(t+1)\|\epsilon\|_2)$;
    Perturb the penalty term according to (2.47);
    Update primal variable via (2.47);
  for $i = 1$ **to** $\mathscr{N}$ **do**
    Broadcast $f_i(t+1)$ to all neighbors $j \in \mathscr{V}_i$;
  for $i = 1$ **to** $\mathscr{N}$ **do**
    Update dual variable according to (2.11);
Output: Upper bound of the total privacy loss β; primal $\{f_i(T)\}_{i=1}^N$, dual $\{\lambda_i(T)\}_{i=1}^N$

2.5.2 MR-ADMM

In odd iterations, we adopt the objective perturbation [24] directly where a random linear term $\epsilon_i(2k-1)^T f_i$ is added to the objective function in (2.14) [3], where $\epsilon_i(2k-1)$ follows the probability density proportional to $\exp\{-\alpha_i(k)\|\epsilon_i(2k-1)\|_2\}$. Consequently the objective function for updating

[2]Only a single iteration is considered in [147] while imposing a privacy constraint. Since we consider the entire iterative process, we don't impose per-iteration privacy constraint but calculate the total privacy loss.

[3]Pure differential privacy was adopted in this work, but the weaker (ϵ, δ)-differential privacy can be applied as well.

the primal variable $f_i(2k-1)$ becomes $L_i^{priv}(2k-1)$ given as follows:

$$\begin{aligned} L_i^{priv}(2k-1) &= O(f_i, D_i) + (2\lambda_i(2k-2) + \epsilon_i(2k-1))^T f_i \\ &\quad + \eta_i(2k-1) \sum_{j \in \mathscr{V}_i} \| \frac{1}{2}(f_i(2k-2) + f_j(2k-2)) - f_i \|_2^2. \end{aligned}$$

To generate this noisy vector $\epsilon_i(2k-1)$, choose the norm from the gamma distribution with shape d and scale $\frac{1}{\alpha_i(k)}$ and the direction uniformly, where d is the dimension of the feature space. Node i's local result (primal variable) is obtained by finding the optimal solution to the private objective function:

$$f_i(2k-1) = \underset{f_i}{\operatorname{argmin}}\ L_i^{priv}(2k-1), \quad i \in \mathscr{N}\ . \tag{2.47}$$

In the $2k$-th iteration, use the stored results $\epsilon_i(2k-1) + \nabla O(f_i(2k-1), D_i)$ and $\eta_i(2k-1) \sum_{j \in \mathscr{V}_i} (f_i(2k-1) - f_j(2k-1))$ to update primal variables, where the latter can be obtained from the dual update in the $(2k-1)$-th update, and the former can be obtained directly from the KKT condition in the $(2k-1)$-th iteration:

$$\epsilon_i(2k-1) + \nabla O(f_i(2k-1), D_i) = -2\lambda_i(2k-2) - \eta_i(2k-1) \sum_{j \in \mathscr{V}_i} (2f_i(2k-1)) - f_i(2k-2) - f_j(2k-2))\ .$$

Then the even update is given by:

$$\begin{aligned} f_i(2k) &= f_i(2k-1) - \frac{1}{2\eta_i(2k-1)V_i + \gamma} \{2\lambda_i(2k-1) + \underbrace{\epsilon_i(2k-1) + \nabla O(f_i(2k-1), D_i)}_{\text{the existing result by KKT}} \\ &\quad \underbrace{+ \eta_i(2k-1) \sum_{j \in \mathscr{V}_i} (f_i(2k-1) - f_j(2k-1))\}}_{\text{the existing result by the previous dual update}}\ . \end{aligned} \tag{2.48}$$

Algorithm 2 shows the complete procedure, where the condition used to generate $\eta_i(1)$ helps to bound the worst-case privacy loss but is not necessary in guaranteeing convergence.

Algorithm 2: Private MR-ADMM

Input: $\{D_i\}_{i=1}^N$, $\{\alpha_i(1), \cdots, \alpha_i(K)\}_{i=1}^N$
Initialize: $\forall i$, generate $f_i(0)$ randomly, $\lambda_i(0) = \mathbf{0}_{d\times 1}$
Parameter: $\forall i$, select $\{\eta_i(2k-1)\}_{k=1}^K$ s.t. $0 < \eta_i(2k-1) \leq \eta_i(2k+1) < +\infty$, $\forall k$ and $\eta_i(1)$ satisfies $2c_1 < \min_i\{\frac{B_i}{C}(\frac{\rho}{N} + 2\eta_i(1)V_i)\}$
for $k = 1$ **to** K **do**
  for $i = 1$ **to** $\mathcal{N}$ **do**
    Generate noise $\epsilon_i(2k-1) \sim \exp(-\alpha_i(k)\|\epsilon\|_2)$;
    Update primal variable $f_i(2k-1)$ via (2.47);
    Broadcast $f_i(2k-1)$ to all neighbors $j \in \mathcal{V}_i$.
  for $i = 1$ **to** $\mathcal{N}$ **do**
    Calculate $\eta_i(2k-1)\sum_{j\in\mathcal{V}_i}(f_i(2k-1) - f_j(2k-1))$;
    Update dual variable $\lambda_i(2k-1)$ via (2.17).
  for $i = 1$ **to** $\mathcal{N}$ **do**
    Use the stored information $\epsilon_i(2k-1) + \nabla O(f_i(2k-1), D_i)$ and $\eta_i(2k-1)\sum_{j\in\mathcal{V}_i}(f_i(2k-1) - f_j(2k-1))$ to update primal variable $f_i(2k)$ via (2.48);
    Keep the dual variable $\lambda_i(2k) = \lambda_i(2k-1)$;
    Broadcast $f_i(2k)$ to all neighbors $j \in \mathcal{V}_i$.
Output: Upper bound of the total privacy loss β; primal $\{f_i(2K)\}_{i=1}^N$, dual $\{\lambda_i(2K)\}_{i=1}^N$

2.6 Privacy Analysis

In this section, we characterize the total privacy loss of private M-ADMM and private MR-ADMM as presented in Algorithms 1 and 2. Similar to the previous section, the results also apply to private R-ADMM by fixing $\eta_i(2k-1) = \eta$, $\forall k$.

As mentioned earlier, Zhang and Zhu [147] only quantifies the privacy loss of a single node in a single iteration, i.e., $\frac{\Pr(f_i(t)\in S_i|D_i)}{\Pr(f_i(t)\in S_i|\hat{D}_i)} \leq \exp(\alpha_i(t))$ holds $\forall t, i$, where $\alpha_i(t)$ is the bound on the privacy loss of node i at iteration t. However, in a distributed and iterative setting, the "output" of the algorithm is not merely the end result, but includes all intermediate results generated and exchanged during the iterative process; an attacker can use all such intermediate results to perform inference. For this reason, we adopt the differential privacy definition proposed in [149] as follows, which bounds the total privacy loss during the entire iterative process.

Definition 1. *Consider a connected network $G(\mathcal{N}, \mathcal{E})$ with a set of nodes $\mathcal{N} = \{1, 2, \cdots, N\}$. Let $f(t) = \{f_i(t)\}_{i=1}^N$ denote the information exchange of all nodes in the t-th iteration. A distributed algorithm is said to satisfy β-differential privacy during T iterations if for any two datasets D_{all} =*

$\cup_i D_i$ *and* $\hat{D}_{all} = \cup_i \hat{D}_i$*, differing in at most one data point, and for any set of possible outputs* S *during* T *iterations, the following holds:*

$$\frac{Pr(\{f(t)\}_{t=0}^{T} \in S \,|D_{all})}{Pr(\{f(t)\}_{t=0}^{T} \in S \,|\hat{D}_{all})} \leq \exp(\beta)$$

The analysis is focused on the regularized empirical risk minimization (ERM) problem for binary classification, while its generalization is discussed in Section 2.8. Let node i's dataset be $D_i = \{(x_i^n, y_i^n) | n = 1, 2, \cdots, B_i\}$, where $x_i^n \in \mathbb{R}^d$ is the feature vector representing the n-th sample belonging to i, $y_i^n \in \{-1, 1\}$ the corresponding label, and B_i the size of D_i. Then the sub-objective function for each node i is defined as follows:

$$O(f_i, D_i) = \frac{C}{B_i} \sum_{n=1}^{B_i} \mathscr{L}(y_i^n f_i^T x_i^n) + \frac{\rho}{N} R(f_i) \,,$$

where $C \leq B_i$ and $\rho > 0$ are constant parameters of the algorithm, the loss function $\mathscr{L}(\cdot)$ measures the accuracy of the classifier, and the regularizer $R(\cdot)$ helps prevent overfitting.

For this binary classification problem, we now state results on the privacy property of the private M-ADMM (Algorithm 1) and private MR-ADMM (Algorithm 2) using Definition 1 above and additional assumptions on $\mathscr{L}(\cdot)$ and $R(\cdot)$ as follows.

Assumption 5. *The loss function* $\mathscr{L}$ *is strictly convex and twice differentiable.* $|\nabla \mathscr{L}| \leq 1$ *and* $0 < \mathscr{L}'' \leq c_1$ *with* c_1 *being a constant.*

Assumption 6. *The regularizer* R *is 1-strongly convex and twice continuously differentiable.*

2.6.1 Private M-ADMM

Theorem 4. *Normalize feature vectors in the training set such that* $\|x_i^n\|_2 \leq 1$ *for all* $i \in \mathscr{N}$ *and* n*. Then the private M-ADMM algorithm (Algorithm 1) satisfies the* β*-differential privacy with*

$$\beta \geq \max_{i \in \mathscr{N}} \{ \sum_{t=1}^{T} \frac{C(1.4c_1 + \alpha_i(t))}{\eta_i(t) V_i B_i} \} \,. \tag{2.49}$$

2.6.2 Private MR-ADMM

Lemma 2. *Consider the private MR-ADMM (Algorithm 2), $\forall k = 1, \cdots K$, assume the total privacy loss up to the $(2k-1)$-th iteration can be bounded by β_{2k-1}, then the total privacy loss up to the $2k$-th iteration can also be bounded by β_{2k-1}. In other words, given the private results in odd iterations, outputting private results in the even iterations does not release more information about the input data.*

Theorem 5. *Normalize feature vectors in the training set such that $\|x_i^n\|_2 \leq 1$ for all $i \in \mathcal{N}$ and n. Then the private MR-ADMM algorithm (Algorithm 2) satisfies the β-differential privacy with*

$$\beta \geq \max_{i \in \mathcal{N}} \{ \sum_{k=1}^{K} \frac{2C}{B_i} (\frac{1.4 c_1}{(\frac{\rho}{N} + 2\eta_i (2k-1) V_i)} + \alpha_i(k)) \} \ . \tag{2.50}$$

2.7 Sample Complexity Analysis.

We next quantify the generalization performance of (non)-private MR-ADMM, the same technique can be applied to M-ADMM. The analysis is focused on the ERM problem defined above and we assume samples from each node i are drawn i.i.d. from a fixed distribution P. The expected loss of node i using classifier $f_i(t)$ at time t is given as $\mathcal{L}(f_i(t)) = \mathbb{E}_{(X,Y)\sim P}(\mathscr{L}(Y f_i(t)^T X))$. Similar to the analysis in [24, 147], we introduce a reference classifier f_{ref} with expected loss $\mathcal{L}(f_{ref})$ and evaluate the generalization performance using the number of samples (B_i) required at each node to achieve $\mathcal{L}(f_i(t)) \leq \mathcal{L}(f_{ref}) + \tau$ with high probability.

2.7.1 Non-Private MR-ADMM

As shown in Section 2.3.3, the sequence of outputs $\{f_i^{non}(2k-1)\}$ from odd iterations in non-private MR-ADMM converges to $f_i^* = f_c^*$ as $k \to \infty$. Therefore, there exists a constant $\Delta_i(k)$ for each node i at the $(2k-1)$-th iteration such that $\mathcal{L}(f_i^{non}(2k-1)) \leq \mathcal{L}(f_c^*) + \Delta_i(k)$. Using the same method as [24, 147], we have the following result.

Theorem 6. *Consider a regularized ERM problem with regularizer $R(f) = \frac{1}{2}\|f\|^2$ and let f_{ref} be a reference classifier for all nodes and $\{f_i^{non}(2k-1)\}$ be a sequence of outputs of non-private*

MR-ADMM in odd iterations (Eqn. (2.16)*). If the number of samples at node i satisfies*

$$B_i \geq w \max_k \Big\{ \frac{\|f_{ref}\|^2 \log(1/\delta)}{(\tau - \Delta_i(k))^2} \Big\}$$

for some constant w, then $f_i^{non}(2k-1)$ *satisfies*

$$\Pr(\mathcal{L}(f_i^{non}(2k-1)) \leq \mathcal{L}(f_{ref}) + \tau) \geq 1 - \delta$$

where $\tau > \Delta_i(k)$, $\forall i, k \in \mathbb{Z}_+$.

As expected, the number of required samples depends on the choice of the reference classifier via its l_2 norm $\|f_{ref}\|^2$, by imposing an upper bound b_{ref} on $\|f_{ref}\|^2$. The result shows that if B_i satisfies $B_i \geq w \max_k \{\frac{b_{ref} \log(1/\delta)}{(\tau - \Delta_i(k))^2}\}$, then the non-private intermediate classifier of each node at odd iterations will have an additional error no more than τ as compared to any classifier with $\|f_{ref}\|^2 \leq b_{ref}$.

2.7.2 Private MR-ADMM

We next present the result on the sample complexity of the private MR-ADMM algorithm. Similar to the analysis of non-private MR-ADMM, we bound the error of the intermediate classifier of each node at odd iterations. Since the algorithm is perturbed with different random noise in different iterations, to better analyze the effect of noise in a single iteration, we adopt a strategy similar to that used in [147], by intentionally fixing the noise in iterations after the targeted iteration. Specifically, $\forall i$, to compare the private $f_i^{priv}(2k-1)$ at the $(2k-1)$-th iteration with reference classifier f_{ref}, we slightly modify Algorithm 2 such that $\forall k' > k$, the added noise is fixed at $\epsilon_i(2k'-1) = \epsilon_i(2k-1)$, which allows us to solely study the effect of $\epsilon_i(2k-1)$. This problem can be formulated as a new MR-ADMM optimization problem where node i's sub-objective function becomes $O^{new}(f_i, D_i) = O(f_i, D_i) + \epsilon_i(2k-1)^T f_i$ and the initialization given by $f_i(0) = f_i(2k-1)$, $\lambda_i(0) = \lambda_i(2k-1)$. Let $\{f_i^{new}(2k-1)\}$ be a sequence of outputs from odd iterations of this new algorithm; it converges to a fixed point f_{new}^* as $k \to \infty$. Therefore, there exists a constant $\Delta_i^{new}(k)$ for each node i at the $(2k-1)$-th iteration such that $\mathcal{L}(f_i^{new}(2k-1)) \leq \mathcal{L}(f_{new}^*) + \Delta_i^{new}(k)$. Using this, we have the following result.

Theorem 7. *Consider a regularized ERM problem with regularizer* $R(f) = \frac{1}{2}\|f\|^2$*, let* f_{ref} *be a reference classifier for all nodes and* $\{f_i^{priv}(2k-1)\}$ *be a sequence of outputs of private MR-ADMM*

in odd iterations. If the number of samples at node i satisfies

$$B_i \geq w \max_k \Big\{ \frac{CN \log(1/\delta)}{\frac{NC(\tau - \Delta_i^{new}(k))^2}{2\|f_{ref}\|^2} - (1+a)\frac{Nd^2}{C(\alpha_i(k))^2}(\log(d/\delta))^2} \Big\}$$

for some constants w and $a > 0$, then $f_i^{priv}(2k-1)$ satisfies

$$\Pr(\mathcal{L}(f_i^{priv}(2k-1)) \leq \mathcal{L}(f_{ref}) + \tau) \geq 1 - 2\delta$$

where $\tau > \Delta_i^{new}(k)$, $\forall i, k \in \mathbb{Z}_+$.

Compared to Theorem 6, we see an additional term imposed by the privacy constraints, i.e., $(1+a)\frac{Nd^2}{C(\alpha_i(k))^2}(\log(d/\delta))^2$. If $\alpha_i(k) \to \infty$, the result reduces to $B_i \geq w \max_k \{\frac{2\|f_{ref}\|^2 \log(1/\delta)}{(\tau - \Delta_i^{new}(k))^2}\}$, the same as given in Theorem 6. The additional term shows that the higher dimension of features, the more injected noise, which would require more samples to achieve the same accuracy.

2.8 Discussion

Improving privacy-accuracy trade-off. We now provide some intuitive explanation as to why the ideas presented in this chapter work. We explored two key ideas to improve the privacy-accuracy tradeoff of a differentially private algorithm. The first is to accomplish the computational task by repeatedly using the already released differentially private outputs. Utilizing differential privacy's immunity to post-processing, this information recycling incurs no additional privacy loss. Since less information is revealed during computation, less perturbation is required to obtain the same privacy guarantee, which then improves the privacy-accuracy tradeoff. The second idea is to improve the the stability/robustness of the algorithm by directly controlling the penalty parameter. This allows the algorithm to accommodate more noise to improve privacy without sacrificing too much accuracy, which improves the privacy-accuracy tradeoff.

Other perturbation methods and privacy analysis tools. While we have primarily used objective perturbation to make an algorithm differentially private and to calculate the privacy loss, it should be noted that this is done as an example to illustrate how MR-ADMM can outperform both R-ADMM and ADMM in the privacy-accuracy tradeoff. Other perturbation methods such as

output perturbation to achieve differential privacy (each node perturbs its primal variable before broadcasting to its neighbors) can be used as well; our conclusion would still hold. This is because our key ideas (revealing less information and making the algorithm more robust/stable to noise via the penalty parameter) are orthogonal to the choice of the perturbation method.

Similarly, in our privacy analysis we have adopted the notion of pure ε-differential privacy to measure privacy. As a result, the bound on the total privacy loss can be fairly large. It is also possible to adopt a weaker notion, the (ε, δ)-differential privacy, to find a tighter bound on privacy loss by allowing the algorithm to violate ε-differential privacy with a small probability δ. In this case, the total privacy loss can be calculated using more advanced composition theorems such as moments accountant [4] and zero-concentrated differential privacy [22]. However, our key ideas (revealing less information and making the algorithm more robust/stable to noise via the penalty parameter) are orthogonal to the choice of the privacy definition and analysis tools used; thus the algorithmic properties will not be affected by such choices and the conclusion remains valid.

Privacy analysis for a broader class of optimizations In Section 2.6, the privacy property of the private algorithms is analyzed for the ERM binary classification problem. This privacy analysis can be extended to more general forms of $O(f_i, D_i)$, such as multi-class settings. There have been extensive studies on the differentially private ERM with convex loss function [132], which can also be adopted for our framework.

2.9 Numerical Experiments

We use the same dataset as [147], i.e., the *Adult* dataset from the UCI Machine Learning Repository [99]. It consists of personal information of around 48,842 individuals, including age, sex, race, education, occupation, income, etc. The goal is to predict whether the annual income of an individual is above \$50,000.

To preprocess the data, we (1) remove all individuals with missing values; (2) convert each categorical attribute (with m categories) to a binary vector of length m; (3) normalize columns (features) such that the maximum value of each column is 1; (4) normalize rows (individuals) such that its l_2 norm is at most 1; and (5) convert labels $\{\geq 50\text{k}, \leq 50\text{k}\}$ to $\{+1, -1\}$. After this preprocessing, the final data includes 45,223 individuals, each represented as a 105-dimensional vector of norm at most 1. We then randomly partition this sample set into a training set (40,000

samples) and a testing set (5,223 samples). The training samples are then evenly distributed across nodes in a network.

We use as loss function the logistic loss $\mathscr{L}(z) = \log(1+\exp(-z))$, with $|\mathscr{L}'| \leq 1$ and $\mathscr{L}'' \leq c_1 = \frac{1}{4}$. The regularizer is $R(f_i) = \frac{1}{2}\|f_i\|_2^2$. We measure the accuracy of the algorithm by the average loss over the training set:

$$L(t) := \frac{1}{N}\sum_{i=1}^{N}\frac{1}{B_i}\sum_{n=1}^{B_i}\mathscr{L}(y_i^n f_i(t)^T x_i^n),$$

and the classification error rate over the testing set $\mathcal{S}_{test}$:

$$E = \frac{\sum_{(x_j,y_j)\in\mathcal{S}_{test}} \mathbf{1}(y_j \neq \hat{y}_j)}{\sum_{(x_j,y_j)\in\mathcal{S}_{test}} 1},$$

where $\hat{y}_j$ is the prediction of sample (x_j, y_j) by using the averaged classifier $\bar{f}(t) = \frac{1}{N}\sum_{i=1}^{N} f_i(t)$, and each $f_i(t)$ is the local classifier(primal variable) of node i after t iterations.

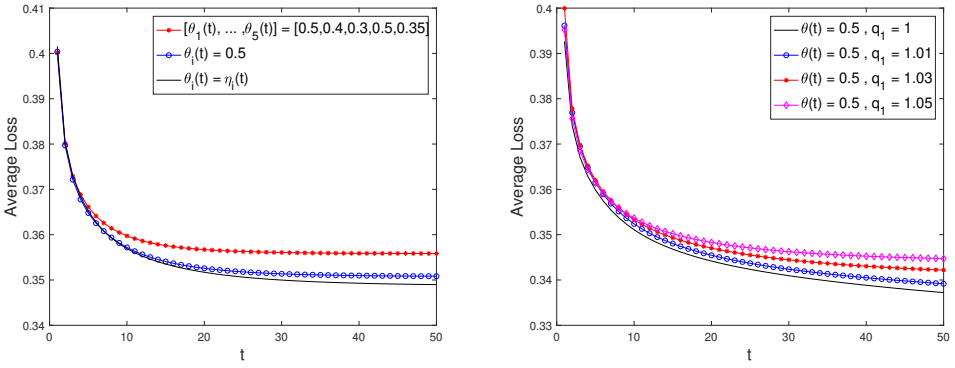
We measure the privacy of an algorithm by the upper bound $P(t)$ given in Theorems 4 and 5. The smaller $L(t)$ and $P(t)$, the higher accuracy and stronger privacy guarantee.

2.9.1 Convergence of Non-Private M-ADMM, R-ADMM & MR-ADMM

We consider a five-node network and assign each node the following private penalty parameters: $\eta_i(t) = \eta_i(1)q_i^{t-1}$ for node i, where $[\eta_1(1), \cdots, \eta_5(1)] = [0.55, 0.65, 0.6, 0.55, 0.6]$ and $[q_1, \cdots, q_5] = [1.01, 1.03, 1.1, 1.2, 1.02]$.

Figure 2.1a shows the convergence of M-ADMM under these parameters while using a fixed dual updating step size $\theta = 0.5$ across all nodes (blue curve). This is consistent with Theorem 1. As mentioned earlier, this step size can also be non-fixed (black) and different (red) for different nodes. In Figure 2.1b we let each node use the same penalty $\eta_i(t) = \eta(t) = 0.5q_1^{t-1}$ and compare the results by increasing $q_1 \geq 1$. We see that increasing penalty slows down the convergence, and larger increase in q_1 slows it down more, which is consistent with Theorem 2.

Figure 2.2a shows the convergence of R-ADMM with different γ and fixed $\eta = 0.5$ for a small network ($N = 5$) and a large network ($N = 20$), both are randomly generated. Due to the linear approximation in even iterations, it's possible to cause an increased average loss as shown in the plot. However, the odd iterations will always compensate this increase; if we only look at the odd iterations, R-ADMM achieves a similar convergence rate as conventional ADMM. γ can also be



(a) Different $\eta_i(t)$ (b) Increasing penalty

Figure 2.1: Convergence properties of M-ADMM.

thought of as an extra penalty parameter for each node in even iterations to punish its update, i.e., the difference between $f_i(2k)$ and $f_i(2k-1)$. Larger γ can result in smaller oscillation between even and odd iterations but will also lower the convergence rate.

Figures 2.2b and 2.2c show the convergence of MR-ADMM with penalty parameters $\eta_i(2k-1)$ increasing at different speed. We see that increasing penalty slows down the convergence, and larger increase in $q_1(i)$ slows it down more. In 2.2b, each node adopts different penalty parameter $\eta_i(2k-1)$ in each iteration while in 2.2c, the same penalty parameter is shared among all the nodes. The convergence is attained in both cases.

2.9.2 Private M-ADMM, R-ADMM & MR-ADMM

The effect of ρ, γ, $\eta_i(2k-1)$ in (M)R-ADMM. We next inspect the accuracy and privacy of the private M-ADMM, R-ADMM and MR-ADMM, and compare it with the private (conventional) ADMM using dual variable perturbation (DVP) [147].

To begin, we first examine the effect of ρ in controlling overfitting. Figure 2.3 shows the classification error rate over the testing set under different ρ, where the classifiers are trained with original ADMM and the algorithm runs for 50 iterations. Since the classification error rate is minimized at $\rho \approx 0.22$, we will use $\rho = 0.22$ in

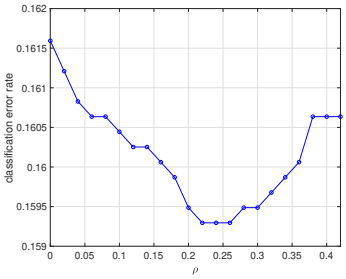
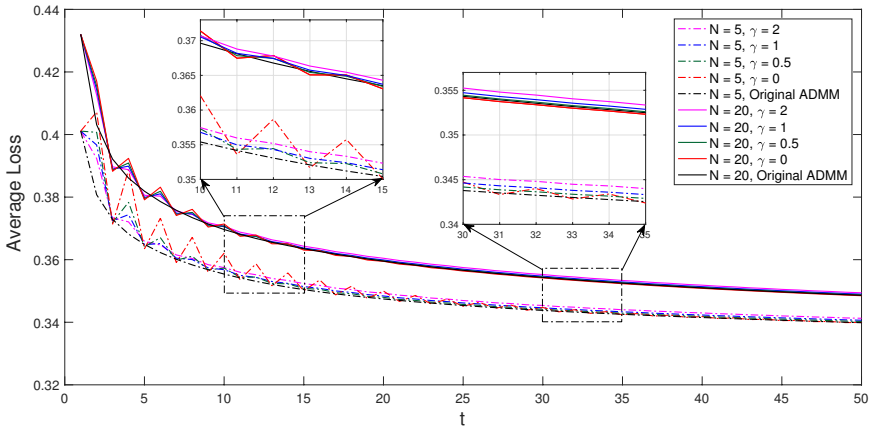
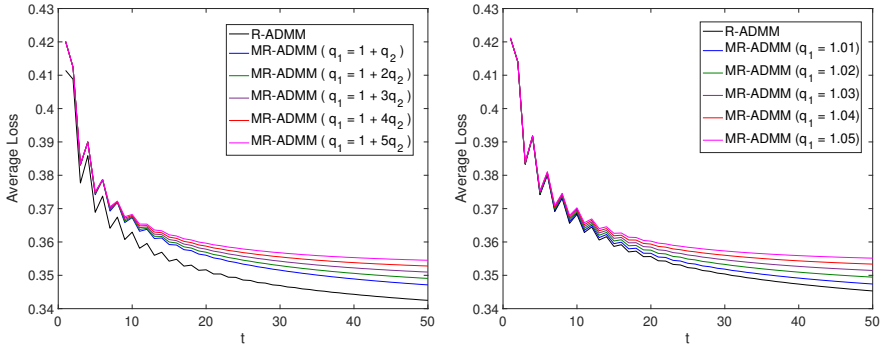


Figure 2.3: The effect of ρ, fixing $C = 1750$.



(a) R-ADMM: $\eta = 0.5$



(b) $\eta_i(2k-1) = \hat{\eta}_i q_1(i)^k$ (c) $\eta_i(2k-1) = q_1^k$

Figure 2.2: Convergence properties of R-ADMM and MR-ADMM: Figure 2.2a illustrates the average loss over iterations of R-ADMM for the network of different sizes under fixed $\eta = 0.5$ and different γ. Dashed (resp. solid) curves represent the performance over a randomly generated small (resp. large) network with $N = 5$ (resp. $N = 20$) nodes. Figures 2.2b and 2.2c illustrate the average loss over iterations of MR-ADMM for a randomly generated network with $N = 5$ nodes. Black curve represents the R-ADMM where $\eta_i(t) = \eta = 1$ is fixed for all nodes and all iterations. Each colored curve represents MR-ADMM with $\eta_i(2k-1)$ increasing over iterations at different speed. In Figure 2.2b, each node i adopts $\eta_i(2k-1) = \eta_i q_1(i)^k$ as penalty parameter in $2k-1$-th iteration, where $[\eta_1, \cdots, \eta_5] = [1, 1.03, 1.02, 0.8, 1.01]$, $q_1 = [q_1(1), \cdots, q_1(5)] = \mathbf{1} + kq_2$ (each $k \in \{1, \cdots, 5\}$ corresponds to one curve in plot) and $q_2 = [q_2(1), \cdots, q_2(5)] = [0.01, 0.005, 0.003, 0.015, 0.01]$. In Figure 2.2c, each node adopts the same penalty parameter $\eta_i(2k-1) = q_1^k$ in odd iterations.

the following experiments.

For simplicity of presentation, in the next set of experiments the penalty $\eta_i(t) = \eta(t)$ in both M-ADMM and MR-ADMM and noise $\alpha_i(k) = \alpha, \forall i, k$. We observe similar results when $\alpha_i(t)$, $\eta_i(t)$ vary from node to node.

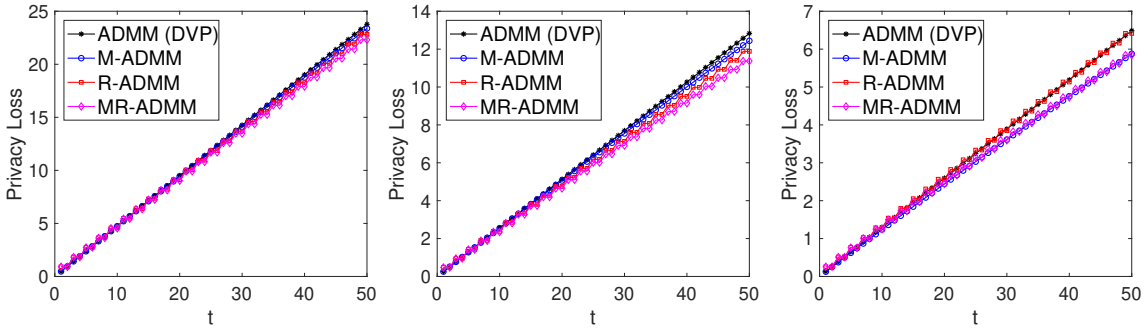
For each parameter setting, we perform 10 independent runs of the algorithm, and record both the mean and the range of their accuracy. Specifically, $L^l(t)$ denotes the average loss over the training dataset in the t-th iteration of the l-th experiment ($1 \leq l \leq 10$). The mean of average loss is given by $L_{mean}(t) = \frac{1}{10}\sum_{l=1}^{10} L^l(t)$ and the range $L_{range}(t) = \max_{1\leq l\leq 10} L^l(t) - \min_{1\leq l\leq 10} L^l(t)$. The larger the range $L_{range}(t)$ the less stable the algorithm, i.e., under the same parameter setting, the difference in performances (convergence curves) of two experiments is larger. In the next few plots, $L_{range}(t)$ is shown as the size of a vertical bar centered at $L_{mean}(t)$. Similarly, let E^l be the classification error rate over the testing set in the l-th experiment, with an average error rate $E_{mean} = \frac{1}{10}\sum_{l=1}^{10} E^l$ and range $E_{range} = \max_{1\leq l\leq 10} E^l - \min_{1\leq l\leq 10} E^l$ shown as the size of a vertical bar centered at E_{mean}. Each parameter setting also has a corresponding upper bound on the privacy loss denoted by $P(t)$.

In the non-private case, γ controls the oscillation between even and odd iterations, as well as the convergence rate. We now examine its effect when MR-ADMM is perturbed. Figure 2.5 shows the average loss over the training set (Figure 2.5a2.5b) and the classification error rate over the testing set (Figure 2.5c) under different $\gamma > 0$, noting that the corresponding privacy loss of these cases are the same under the same α. It shows that varying γ (within a certain range) does not effect performance significantly. For the next set of experiments, we fix $\gamma = 0.5$.

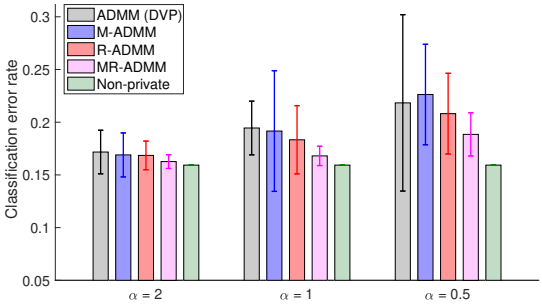
The effect of $\eta_i(2k-1)$ on the performance of private MR-ADMM is illustrated in Figure 2.6, where the pair Figure 2.6a, 2.6c is for the case when noise parameter is $\alpha = 2$ (low privacy requirement) and the pair Figure 2.6b, 2.6d is for the case when $\alpha = 1$ (high privacy requirement). Although increasing $\eta_i(2k-1)$ over time can decrease the convergence rate of non-private MR-ADMM (Figures 2.2b and 2.2c), it helps to stabilize the algorithm when MR-ADMM is perturbed and can improve the accuracy while maintain the privacy guarantee. Moreover, the improvement is more significant when algorithm is under higher perturbation (high privacy requirement) and when $\eta_i(2k-1)$ increases faster (within a range).

Performance comparison among different algorithms. Our last set of experiments is conducted to compare the performance of different algorithms with results illustrated in Figures 2.7 and 2.4. The noise parameters of both MR-ADMM and R-ADMM are set as α shown in the plots, and the

noise parameters of conventional ADMM and M-ADMM are chosen respectively such that they have approximately the same total privacy loss bounds. We set $\eta_i(2k-1) = 1.04^k$ in MR-ADMM. We see that both private R-ADMM (red) and private MR-ADMM (magenta) outperform private ADMM (black) and M-ADMM (blue) with higher accuracy and lower privacy loss. In particular, the private MR-ADMM (magenta) has the highest accuracy with the lowest privacy loss among all algorithms; the improvement is more significant with smaller total privacy loss. This improvement is also illustrated by the classification error rate over the testing set in Figure 2.4d.

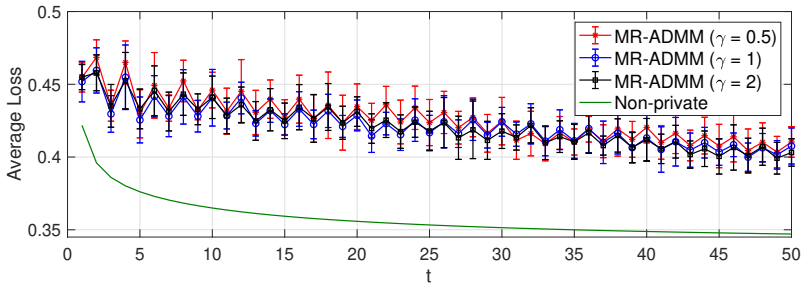


(a) $\alpha = 2$ (b) $\alpha = 1$ (c) $\alpha = 0.5$

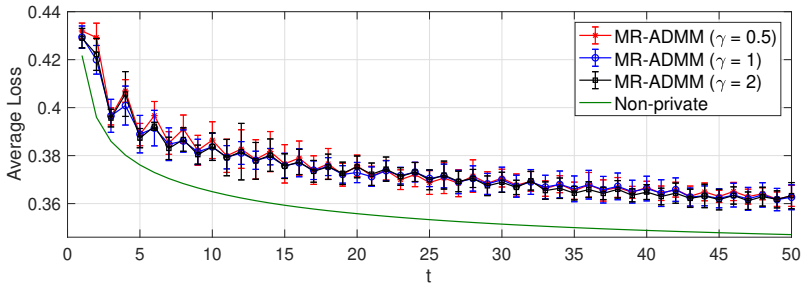


(d) Classification error rate comparison

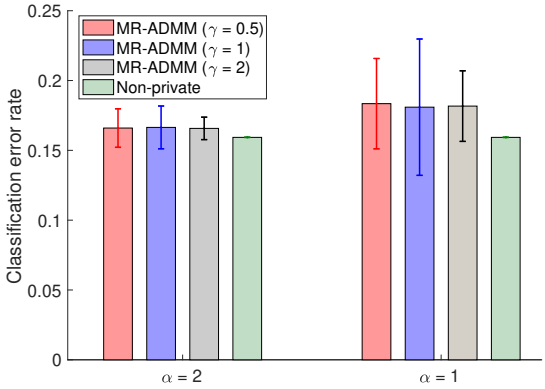
Figure 2.4: Performance comparison: Figures 2.4a, 2.4b and 2.4c illustrate the upper bound of their privacy loss and the corresponding classification error rates are shown in Figure 2.4d.



(a) Accuracy comparison for different γ ($\alpha = 1$)

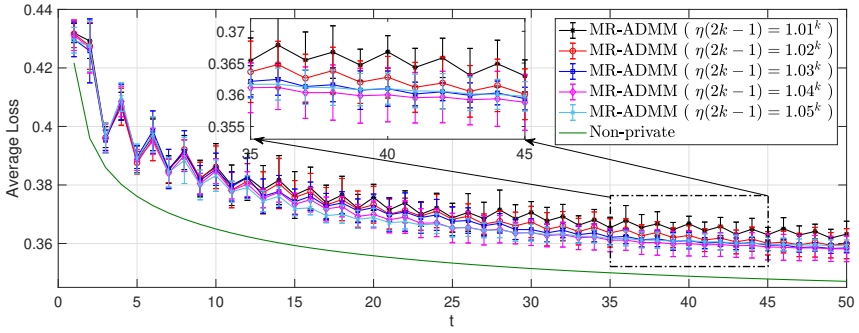


(b) Accuracy comparison for different γ ($\alpha = 2$)

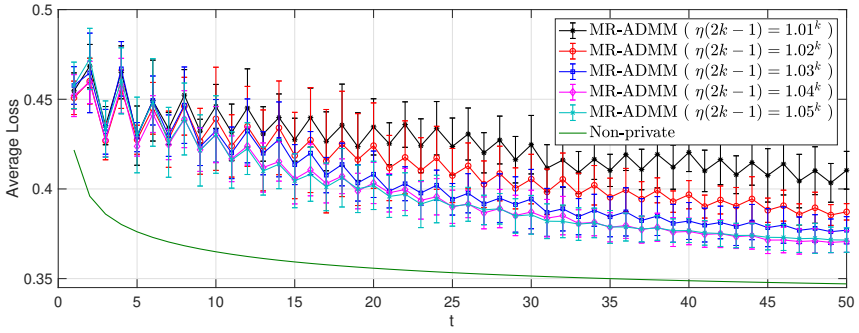


(c) Classification error rate comparison

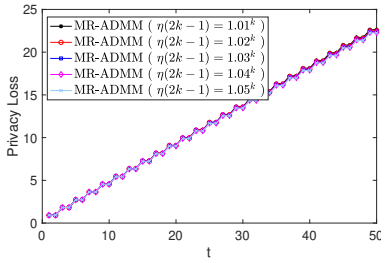
Figure 2.5: The effect of γ on the performance of MR-ADMM, fixing $\eta_i(2k-1) = 1.01^k$: in Figures 2.5a and 2.5b, green curves represent the non-private conventional ADMM while other curves represent the private MR-ADMM with different γ and each of them illustrates the overall result summarized from 10 independent runs of experiments under the same parameter. The corresponding classification error rates are shown in Figure 2.5c. It shows that varying γ within a certain range doesn't effect the performance significantly.



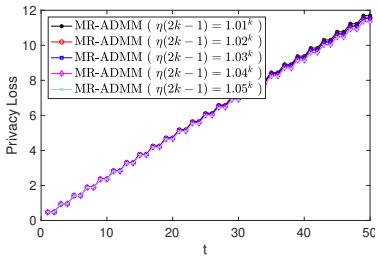
(a) Accuracy comparison for different $\eta(2k-1)$ ($\alpha = 2$)



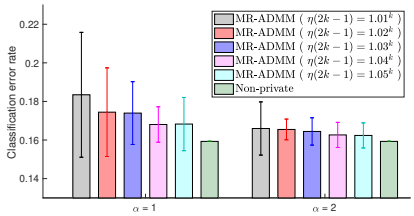
(b) Accuracy comparison for different $\eta(2k-1)$ ($\alpha = 1$)



(c) Privacy comparison ($\alpha = 2$)

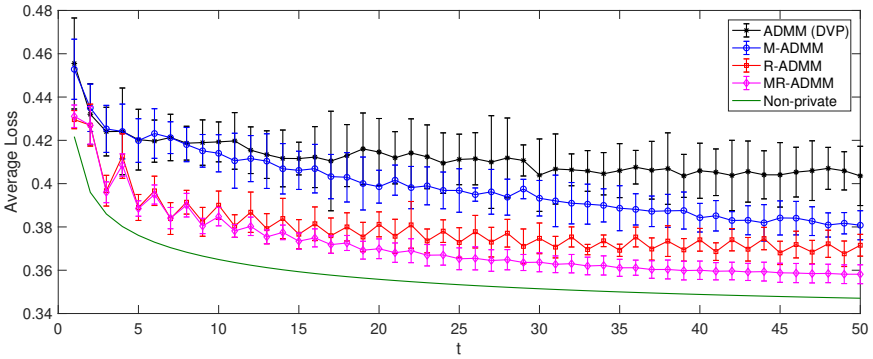


(d) Privacy comparison ($\alpha = 1$)

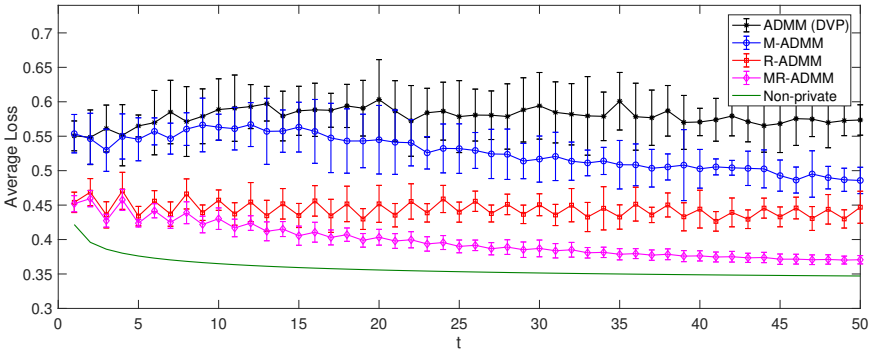


(e) Classification error rate

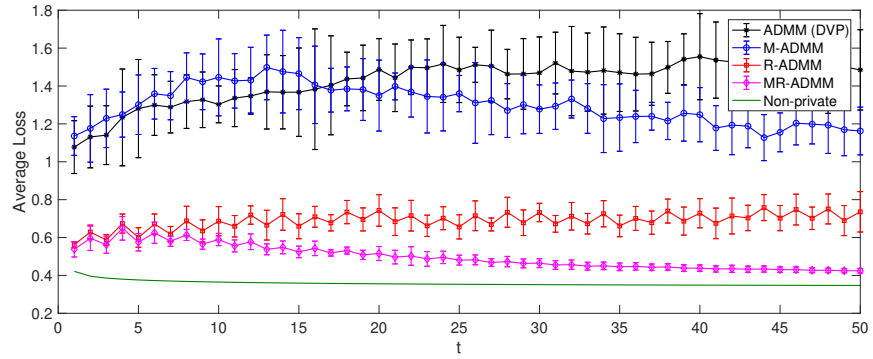
Figure 2.6: The effect of $\eta_i(2k-1)$ on the performance of MR-ADMM, fixing $\gamma = 0.5$: in Figures 2.6a and 2.6b, green curves represent the non-private conventional ADMM while other curves represent the private MR-ADMM with different $\eta_i(2k-1) = q_1^k$ ($q_1 = 1.01, 1.02, 1.03, 1.04, 1.05$) and each of them illustrates the overall result summarized from 10 independent runs of experiments under the same parameter. Figures 2.6c and 2.6d illustrate the upper bound of their privacy loss and the corresponding classification error rates are shown in Figure 2.6e.



(a) Accuracy comparison ($\alpha = 2$)



(b) Accuracy comparison ($\alpha = 1$)



(c) Accuracy comparison ($\alpha = 0.5$)

Figure 2.7: Performance comparison: in Figures 2.7a, 2.7b and 2.7c, green curves represent the non-private conventional ADMM while other curves represent different private algorithms and each of them illustrates the overall result summarized from 10 independent runs of experiments under the same parameter. M-ADMM (blue) and MR-ADMM (magenta) adopt the varied penalty parameter while ADMM (black) and R-ADMM (red) adopt the fixed $\eta_i(t) = \eta = 1$.

CHAPTER 3

Real-Time Release of Sequential Data with Differential Privacy

3.1 Introduction

In Chapter 2, we explored two ideas that can be leveraged to improve an algorithm's privacy-accuracy tradeoff: (1) reuse intermediate computations to reduce information leakage; (2) improve the robustness to accommodate more noise. These two ideas are not limited to distributed learning but are applicable to many other applications. In this chapter, we consider sequential computation and illustrate how can we leverage idea (1) when designing an algorithm for privately releasing the sequential data in real-time.

The collection and analysis of sequential data are crucial for many applications, such as monitoring web browsing behavior, analyzing daily physical activities recorded by wearable sensors, and so on. Privacy concerns arise when data is shared with third parties, a common occurrence. Toward this end, differential privacy [35] has been widely used to provide a strong privacy guarantee; it is generally achieved by disclosing a noisy version of the underlying data so that changes in the data can be effectively obscured.

To achieve differential privacy in sharing sequential data, a simple approach is to add independent noise to the data at each time instant (Figure 3.1a). This is problematic because of the temporal correlation in the data (see Section 3.3). A number of studies have attempted to address this issue. For example, [121] applies Discrete Fourier Transform (DFT) of the sequence and release a private version generated using inverse DFT with the perturbed DFT coefficients; [133] proposes a correlated perturbation mechanism where the correlated noise is generated based on the autocorrelation of the original sequence; [80] decomposes the sequence into disjoint groups of

similar data, and uses the noisy averages of these groups to reconstruct the original sequence; [138] constructs a Hidden Markov Model (HMM) from the independent-noise-added data sequence, and releases the sequence inferred from the HMM; method proposed in [44] first reconstructs the non-sampled data from perturbed sampled points and then solves a convex optimization to improve accuracy. However, all of the above studies rely on the availability of the entire sequence, so can only be applied offline as post-processing methods. [41] are the closest to our work, where the sequence is adaptively sampled first; Kalman/particle filters are then used to estimate non-sampled data based on the perturbed sampled data. However, it requires a priori knowledge of the correlation of the sequence.

In this chapter we start from sequential data that can be modeled by first-order autoregressive (AR(1)) processes. We consider Gaussian AR(1) process and Binomial AR(1) process as examples but the idea can be generalized to all (weakly) *stationary* processes. Leveraging time-invariant statistical properties of stationary process, proposed approach in each time step estimates the unreleased, future data from that already released, using correlation learned over time and not required a priori. This estimate is then used, in conjunction with the actual data observed in the next time step, to drive the generation of the noisy, released version of the data (Figure 3.1b). Both theoretical analysis and empirical results show that our approach can release a sequence of high accuracy with less privacy loss.

Our main findings and contributions are as follows.

1. We develop a method for releasing data sequence in real time with differential privacy guarantee (Sections 3.4).
2. We conduct privacy and accuracy analyses to theoretically quantify the total privacy loss (Section 3.5) and error (Section 3.6).
3. For sequences following Gaussian AR(1) processes, we show erotically that the proposed method can *strictly* outperform the baseline method (Section 3.6).
4. We conduct experiments on real-world data to show the effectiveness of our method.

The rest of the chapter is organized as follows. Section 3.2 presents background and preliminaries. Section 3.3 introduces the baseline approach and its issues. Our approach is presented and analyzed in Sections 3.4, 3.5 and 3.6. Section 3.7 presents Discussion. Experiments are presented in Section 3.8. All proofs are presented in Appendix B.

3.2 Preliminaries

Consider a time-varying sequence $\{Z_t\}_{t=1}^T$, where $Z_t \in \mathbb{R}$ corresponds to a query over a private dataset D_t at time $t \in \mathbb{N}$, i.e., $Z_t = \mathcal{Q}(D_t)$. Dataset $D_t = \{d_t^i\}_{i=1}^N$ consists of data from N individuals ($N \geq 1$) where d_t^i is the data of i^{th} individual at time step t. Then $d_{1:T}^i = \{d_t^i\}_{t=1}^T$ is the data of i^{th} individual over T time steps and $D = \{d_{1:T}^i\}_{i=1}^N$ includes sequences of N individuals over T time steps.

We assume $\{Z_t\}_{t=1}^T$ can be modeled as a first-order autoregressive (AR(1)) process [135], where the value at each time depends linearly on the value at immediate preceding time step, but we will see the approach can be generalized to any (weakly) *stationary* process. The goal is to disclose/release this data in real time with privacy guarantees for each individual at all times. We denote by $\{X_t\}_{t=1}^T$ the released sequence. Notationally, we will use X to denote a random variable with probability distribution $\mathscr{F}_X(\cdot)$, x its realization and $\hat{X}(y)$ the estimate of X given observation $Y = y$; finally, $X_{1:t} := \{X_i\}_{i=1}^t$.

3.2.1 First-Order Autoregressive Process

AR(1) processes are commonly used for modeling a time-series, among which Gaussian AR(1) process is one type that is widely used in various domains.

Definition 2 ((Gaussian AR(1) process))**.** *$Z_{1:T}$ is a Gaussian AR(1) process [135] if:*

$$Z_t = \alpha + \rho Z_{t-1} + U_t,\ t \geq 1 \tag{3.1}$$

where $U_t \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_u^2)$, $Z_0 \sim \mathcal{N}(\mu, \sigma_z^2)$ and σ_u^2, α, ρ are constants. If $|\rho| < 1$, then $\{Z_t\}_{t=1}^T$ is a stationary Markov process with the following properties: (1) $Z_t \sim \mathcal{N}(\mu, \sigma_z^2)$ with $\mu = \frac{\alpha}{1-\rho}$ and $\sigma_z^2 = \frac{\sigma_u^2}{1-\rho^2}$; (2) its autocorrelation function is given by $Corr(Z_t Z_{t-\tau}) = Corr(\tau) = \rho^{|\tau|}$.

In addition to Gaussian AR(1) process, we also consider Binomial AR(1) process defined as follows.

Definition 3 (**Binomial AR(1) process**)**.** *Let $\pi \in (0,1)$ and $\rho \in [\max(-\frac{\pi}{1-\pi}, -\frac{1-\pi}{\pi}), 1]$. Define $\beta = \pi(1-\rho)$, $\alpha = \beta + \rho$, and fix $n \in \mathbb{N}$. Then $Z_{1:T}$ is a binomial AR(1) process [109] if:*

$$Z_t = \alpha \circ Z_{t-1} + \beta \circ (n - Z_{t-1}),\ t \geq 1 \tag{3.2}$$

where $Z_0 \sim Binomial(n,\pi)$ and "$\circ$" is called the thinning operator defined as $a \circ Z_{t-1} = \sum_{i=1}^{Z_{t-1}} Y_{i,t-1}$, where $Y_{i,t-1}, i = 1, \cdots, Z_{t-1}$ are i.i.d Bernoulli random variables with $Pr(Y_{i,t-1} = 1) = a$, and all thinnings are independent of each other. Binomial AR(1) is also a stationary Markov process with the following properties: (1) $Z_t \sim Binomial(n,\pi)$; (2) its autocorrelation is $Corr(Z_t Z_{t-\tau}) = Corr(\tau) = \rho^{|\tau|}$.

Binomial AR(1) is typically used for modeling integer-valued counts sequences. Consider n independent entities, each of which can be either in state "1" or state "0". Then Z_t can be interpreted as the number of entities in state "1" at time t. Eqn. (3.2) implies that this "1"-entity count (Z_t) can be given by the number of "1"-entities in the previous time instant that didn't change state ($\alpha \circ Z_{t-1}$) plus the number of "0"-entities in the previous time instant that changed to state "1" ($\beta \circ (n - Z_{t-1})$); here α, β can be interpreted as the respective transition probabilities. Binomial AR(1) has been used to model many real-world scenarios such as counts of computer log-ins and log-outs [136], daily counts of occupied rooms in a hotel, etc.

3.2.2 Differential Privacy

In this chapter, we adopt notion of (ϵ, δ)- differential privacy. We consider the setting where each individual's data is of a sequential nature and a query $\mathcal{Q}$ over N individuals is released over T time steps as it is generated. Within this context, $x_{1:T} = \mathcal{A}(z_{1:T}) = \mathcal{A}(\{\mathcal{Q}(D_t)\}_{t=1}^{T})$ and a randomized algorithm $\mathcal{A}(\cdot)$ is (ϵ, δ)-differentially private if $\mathscr{F}_{X_{1:T}|Z_{1:T}}(x_{1:T}|z_{1:T}) \leq \exp(\epsilon) \cdot \mathscr{F}_{X_{1:T}|Z_{1:T}}(x_{1:T}|\widehat{z}_{1:T}) + \delta$ holds for any possible $x_{1:T}$ and any pairs of $z_{1:T}, \widehat{z}_{1:T}$ generated from D, $\widehat{D}$, where D, $\widehat{D}$ are datasets differing in at most one individual's sequence.[1] It suggests that the released sequence $x_{1:T}$ should be relatively insensitive to the change of one individual's sequential data, thereby preventing the meaningful inference about each individual from observing $x_{1:T}$.

Definition 4 (**Sensitivity of query $\mathcal{Q}$ at time t**). *Consider a query $\mathcal{Q} : \mathcal{D} \to \mathbb{R}$ taking a dataset as input, the sensitivity of $\mathcal{Q}$ at t is defined as: $\Delta\mathcal{Q}_t = \sup_{\forall D_t \sim \widehat{D}_t} |\mathcal{Q}(D_t) - \mathcal{Q}(\widehat{D}_t)|$, where $D_t, \widehat{D}_t \in \mathcal{D}$ are two datasets at t different in at most one individual's data.*

Since $Z_t = \mathcal{Q}(D_t)$, $\Delta\mathcal{Q}_t$ quantifies the maximum impact of an individual on Z_t. In the rest of chapter, unless explicitly stated, we consider scenarios where $\Delta\mathcal{Q}_t$ does not change over time and simplify notation $\Delta\mathcal{Q}_t = \Delta$. If $d_t^i \in \{0, 1\}$, $\forall t$ and $\mathcal{Q}(D_t) = \sum_{i=1}^{N} d_t^i$ is count query (e.g., daily count of patients), then $\Delta = 1$.

[1] If we express D in matrix form, i.e., $D \in \mathbb{R}^{N \times T}$ with $D_{it} = d_t^i$, then D and $\widehat{D}$ are different in at most one row.

3.2.3 Minimum Mean Squared Error Estimate

The minimum mean squared error (MMSE) estimate of a random variable X given observation $Y = y$ is $\hat{X}(y) = \text{argmin}_h \, \mathbb{E}_X((X - h(Y))^2 | Y = y) = \mathbb{E}(X|Y = y)$. If $h(\cdot)$ is constrained to be linear, i.e., $h(Y) = k_1 Y + k_2$, then the corresponding minimization leads to the linear MMSE (LMMSE) estimate and is given by $\hat{X}(y) = \rho_{XY} \frac{\sigma_X}{\sigma_Y}(y - \mathbb{E}(Y)) + \mathbb{E}(X)$ with the mean squared error (MSE) = $(1 - \rho_{XY}^2)\sigma_X^2$, where ρ_{XY} is the correlation coefficient of X and Y, σ_X^2, σ_Y^2 the variance of X, Y respectively. Using these properties, we have the following result.

Proposition 1. *Consider a Gaussian AR(1) process $Z_{1:T}$ defined by* (3.1)*, the MMSE estimate of Z_{t+1} given $Z_t = z_t$ is $\hat{Z}_{t+1}(z_t) = \mu(1 - \rho) + \rho z_t$, with MSE $\sigma_z^2(1 - \rho^2)$. If we use a perturbed $X_i = Z_i + N_i$, $i \in \{1, \cdots, t\}$ to estimate Z_{t+1}, where $N_i \sim \mathcal{N}(0, \sigma_n^2)$ is the added noise, then the MMSE estimate of Z_{t+1} given $X_i = x_i$ is*

$$\hat{Z}_{t+1}(x_i) = \mu(1 - \rho^{t+1-i} \frac{\sigma_z^2}{\sigma_z^2 + \sigma_n^2}) + \rho^{t+1-i} \frac{\sigma_z^2}{\sigma_z^2 + \sigma_n^2} x_i$$

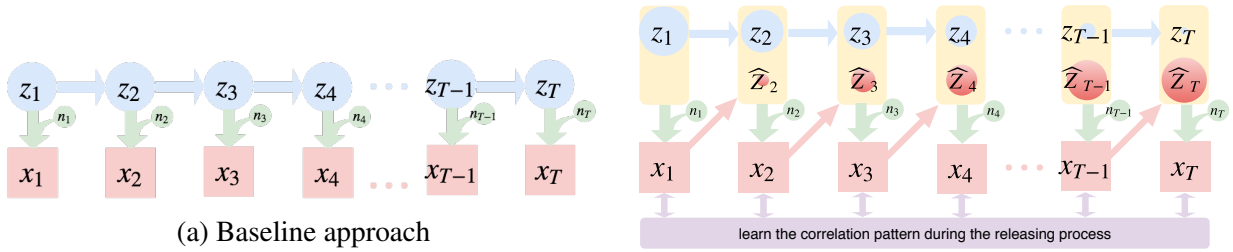
Proposition 2. *Consider a Binomial AR(1) process $Z_{1:T}$ defined by* (3.2)*, the MMSE estimate of Z_{t+1} given $Z_t = z_t$ is $\hat{Z}_{t+1}(z_t) = \rho z_t + n\pi(1 - \rho)$.If we use a perturbed $X_i = Z_i + N_i$, $i \in \{1, \cdots, t\}$ with $Var(N_t) = \frac{m}{2}$ to estimate Z_{t+1}, then the LMMSE estimate of Z_{t+1} given $X_i = x_i$ is*

$$\hat{Z}_{t+1}(x_i) \quad = \quad n\pi(1 - \rho^{t+1-i} \frac{n\pi(1-\pi)}{n\pi(1-\pi) + \frac{m}{2}}) + \rho^{t+1-i} \frac{n\pi(1-\pi)}{n\pi(1-\pi) + \frac{m}{2}} x_i$$

Note that for Gaussian AR(1) processes, both MMSE estimates $\hat{Z}_{t+1}(z_t)$, $\hat{Z}_{t+1}(x_i)$ are linear. For Binomial AR(1), the MMSE estimate $\hat{Z}_{t+1}(z_t)$ is also linear, which may not hold for other AR(1) processes. However, due to the simple form of linear MMSE estimate and its applicability to more general random processes, we will solely focus on LMMSE estimates in this study.

3.3 Baseline Approach

The baseline approach (Figure 3.1a) provides differential privacy for a sequence $z_{1:T}$ by perturbing each z_t directly: $x_t = z_t + perturbation$. However, with this approach it is difficult to obtain a good privacy-accuracy tradeoff for sequences spanning a long time horizon T.



(a) Baseline approach

(b) Proposed approach

Figure 3.1: Comparison of two data release methods:$\{z_t\}_{t=1}^T$ is the true sequence, $\{x_t\}_{t=1}^T$ the released private sequence, $\widehat{Z}_t$ the estimate of z_t learned from x_{t-1}, and $\{n_t\}_{t=1}^T$ the added noise.

The upper bound of the total privacy loss, ϵ_T, can be characterized as a log-likelihood ratio of the released output under two sequences, which can be decomposed as follows:

$$\log \frac{\mathscr{F}_{X_{1:T}|Z_{1:T}}(x_{1:T}|z_{1:T})}{\mathscr{F}_{X_{1:T}|Z_{1:T}}(x_{1:T}|\hat{z}_{1:T})} = \sum_{t=1}^{T} \log \frac{\mathscr{F}_{X_t|Z_t}(x_t|z_t)}{\mathscr{F}_{X_t|Z_t}(x_t|\hat{z}_t)},$$

where the term $\log \frac{\mathscr{F}_{X_t|Z_t}(x_t|z_t)}{\mathscr{F}_{X_t|Z_t}(x_t|\hat{z}_t)}$ bounds the privacy loss at time t. As the total privacy loss is accumulated over T time steps, balancing the privacy-accuracy tradeoff becomes more and more difficult as T increases. As long as the variance of perturbation is finite, as $T \to \infty$, ϵ_T inevitably approaches infinity.

We therefore proposal a method that can (i) improve the privacy-accuracy tradeoff significantly, and (ii) bound the total privacy loss over an infinite horizon when the variance of perturbation is finite.

3.4 The Proposed Approach

In our proposed method, data point x_t at time step t is released based on the previous released data x_{t-1} and its true value z_t (shown in Figure 3.1b).

The idea behind our approach is based on two observations: (1) Since x_{t-1} is correlated with z_t through z_{t-1}, we can use x_{t-1} to obtain an estimate[2] of z_t, denoted by $\hat{Z}_t(x_{t-1})$, and release (perturbed

[2]This estimate can be obtained with or without the knowledge of the statistics of the AR(1) process; in the absence of such knowledge one can employ a separate procedure to first estimate the statistics as detailed later in this section.

version of) $\hat{Z}_t(x_{t-1})$ instead of z_t. (2) Since differential privacy is immune to post-processing [37], using x_{t-1} to estimate z_t does not introduce additional privacy loss. Thus, technically we can release an initial x_1 (perturbed version of z_1), followed by the sequence $x_t = \hat{Z}_t(x_{t-1}), t > 1$. However, doing so will lead to a fairly inaccurate released sequence compared to the original, for while the privacy loss does not accumulate over time, the estimation error does. To balance the competing needs of accuracy (having the released sequence resemble the true sequence) and privacy, one must calibrate the released version using the true values.

There are different ways to calibrate the released sequence. In this study, we shall examine the use of the convex combination $(1-w_t)\hat{Z}_t(x_{t-1}) + w_t z_t$, and the perturbed version of this as the released x_t. Examples of other approaches to calibrating released sequences are discussed in Section 3.7. The weight parameter w_t serves four purposes:

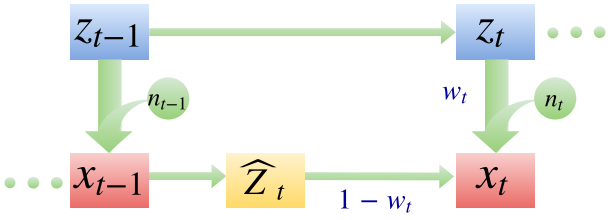


Figure 3.2: A two-step illustration of the proposed method: adding noise n_t to the convex combination of estimate $\hat{Z}_t(x_{t-1})$ and true value z_t gives the released x_t.

(1) In addition to the perturbation σ_n^2, w_t can also be tuned to better balance the privacy-accuracy tradeoff: larger w_t results in a more accurate but less private sequence. In contrast, σ_n^2 is the only means of controlling this tradeoff in the baseline method.

(2) If MSE is the measure of accuracy, then w_t can also be used to balance the bias-variance tradeoff. For a deterministic sequence $z_{1:T}$ with estimator X_t at t, $Bias(X_t) = \mathbb{E}(X_t) - z_t$ and $MSE(X_t) = \mathbb{E}((X_t - z_t)^2)$. The bias and variance can be controlled jointly by adjusting w_t and σ_n^2, which can result in smaller MSE as $MSE = Variance + Bias^2$. In contrast, $x_{1:T}$ is always unbiased in the baseline method and $MSE = Variance$ always holds.

(3) If we keep w_t private, then the method can prevent certain attackers from knowing the detail of the perturbation mechanism, resulting in stronger protection (Section 3.7).

(4) By adjusting w_t, it is possible to release the sequence spanning an infinite horizon with bounded total privacy loss (Section 3.5).

3.4.1 Estimate of Z_t with Learned Correlation

We can estimate the true value Z_t from x_{t-1} using the LMMSE estimate $\hat{Z}_t(x_{t-1})$ given in Section 3.2. However, it requires the knowledge of mean μ, variance σ^2 and autocorrelation ρ of $Z_{1:T}$, which

Algorithm 3: `Est`

Input : $x_{1:t-1}$, $\mathrm{Var}(N_t)$

$\hat{\mu} = \frac{1}{t-1}\sum_{i=1}^{t-1} x_i$

$\hat{\sigma}^2 = \max\{\frac{1}{t-2}\sum_{i=1}^{t-1}(x_i-\hat{\mu})^2 - \mathrm{Var}(N_t), 0\}$

$\hat{\rho} = \frac{\sum_{i=1}^{t-2}(x_i-\hat{\mu})(x_{i+1}-\hat{\mu})}{\sum_{i=1}^{t-2}(x_i-\hat{\mu})^2} + \frac{1}{t-1}\mathbf{1}\{\rho^{true} > 0\}$[3]

Output : $\hat{\rho}, \hat{\mu}, \hat{\sigma}^2$

may be unknown in reality and should be estimated. To avoid revealing more information about $z_{1:T}$, this estimate is obtained using only the released $x_{1:T}$, as shown in Algorithm 3, where both $\hat{\mu}$ and $\hat{\sigma}^2$ are unbiased, and $\hat{\rho}$ is adopted from [69][4].

Release x_t with Estimate $\hat{Z}_t(x_{t-1})$ and True Value z_t. Given the estimated parameters $\hat{\mu}, \hat{\sigma}^2$ and $\hat{\rho}$, using results presented in Section 3.2, the LMMSE estimate $\hat{Z}_t(x_{t-1})$ can be approximated as

$$\hat{Z}_t(x_{t-1}) \quad = \quad \hat{\mu}_{t-1}(1-\hat{\rho}_{t-1}\frac{\hat{\sigma}^2_{t-1}}{\hat{\sigma}^2_{t-1}+\mathrm{Var}(N_t)}) + \hat{\rho}_{t-1}\frac{\hat{\sigma}^2_{t-1}}{\hat{\sigma}^2_{t-1}+\mathrm{Var}(N_t)}x_{t-1} \ .$$

Take the convex combination of estimate $\hat{Z}_t(x_{t-1})$ and true value z_t with private weight w_t, and release:

$$x_t = (1-w_t)\hat{Z}_t(x_{t-1}) + w_t z_t + \textit{perturbation} \ .$$

3.4.2 Privacy Mechanism

The perturbation term in the released data adds privacy protection. Existing literature provides methods on how to generate them. We shall adopt Gaussian mechanism [37] and bound the privacy loss in terms of perturbation.

Lemma 3 (Gaussian Mechanism)**.** *Consider query $\mathcal{Q}: \mathcal{D} \to \mathbb{R}$ with sensitivity $\Delta\mathcal{Q}$, and the Gaussian mechanism $\mathcal{G}(d) = \mathcal{Q}(d) + N$ which adds zero-mean Gaussian noise N with variance σ^2 to the output. If $\sigma \geq \frac{\Delta\mathcal{Q}\sqrt{2\log(1.25/\delta)}}{\epsilon}$ for $\epsilon, \delta \in (0,1)$, then it satisfies (ϵ,δ)-differential privacy.*

Definition 5 (Binomial noise)**.** *We call random variable N the binomial noise if it is zero mean and follows the shifted binomial distribution, i.e., $N + m \sim$ Binomial $(2m, \frac{1}{2})$, whose probability mass*

[4]Extra term $\frac{1}{t-1}$ is used to correct the negative bias if there is prior knowledge of positive autocorrelation $\rho^{true} > 0$.

Algorithm 4: Sequential Data Release Algorithm

Input : Sensitivity of query Δ, $\{\mathrm{Var}(N_t)\}_t$
for $t = 1, 2, \cdots, T$ **do**
  Input : true state z_t, weight w_t
  if $t \leq 2$ **then**
    $w_t = 1$;
    Release : $x_t = z_t + n_t$.
  else
    $\hat{\rho}_{t-1}, \hat{\mu}_{t-1}, \hat{\sigma}^2_{t-1} = \mathtt{Est}(x_{1:t-1}, \mathrm{Var}(N_t))$;
    $r_t = \hat{\rho}_{t-1} \frac{\hat{\sigma}^2_{t-1}}{\hat{\sigma}^2_{t-1} + \mathrm{Var}(N_t)}$;
    Release : $x_t = (1 - w_t)(\hat{\mu}_{t-1}(1 - r_t) + r_t x_{t-1}) + w_t z_t + n_t$
Output : privacy parameter (ϵ_T, δ_T)

function (PMF) is

$$Pr(N = k) = \binom{2m}{k+m} \frac{1}{2^{2m}},\ k \in \{-m, \cdots, m-1, m\},$$

with a variance $\frac{m}{2}$.

Lemma 4 (Binomial Mechanism)**.** *Consider a query* $\mathcal{Q} : \mathcal{D} \to \mathbb{Z}$ *that takes data* $d \in \mathcal{D}$ *as input and outputs an integer. The Binomial mechanism* $\mathcal{B}(d) = \mathcal{Q}(d) + N$ *adds binomial noise* N *with variance* $\frac{m}{2}$ *to the output. If* $1 \leq \Delta\mathcal{Q} + \frac{2m+1}{\exp(\frac{\epsilon}{\Delta\mathcal{Q}})+1} \leq m + 1$ *for* $\epsilon > 0$*, then the following holds:*

(i) $\forall\ \epsilon > 0$*, it satisfies* (ϵ, δ)*-differential privacy with:*

$$\delta = \exp(-\frac{1}{m}(m - \Delta\mathcal{Q} + 1 - \frac{2m+1}{\exp(\frac{\epsilon}{\Delta\mathcal{Q}}) + 1})^2).$$

(ii) $\forall\ \delta \in (0, 1)$*, it satisfies* (ϵ, δ)*-differential privacy with:*

$$\epsilon = \Delta\mathcal{Q} \log(\frac{2m+1}{m - \Delta\mathcal{Q} + 1 - \sqrt{m \log \frac{1}{\delta}}} - 1).$$

Note that Binomial mechanism above is a generalization (for arbitrary sensitivity $\Delta\mathcal{Q}$) to the version (for the case $\Delta\mathcal{Q} = 1$) first proposed in [36]. This is an approximation of the Gaussian mechanism; it has a much looser bound compared to the latter and more noise is needed to ensure a same level of privacy, which is consistent with the conclusion in [36]. However, the Gaussian mechanism only works when $\epsilon < 1$, while our Binomial mechanism does not have this restriction

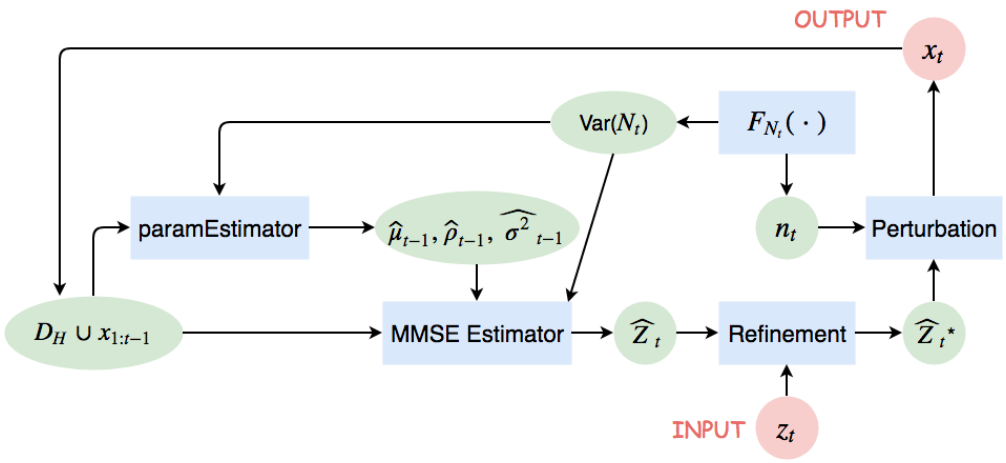


Figure 3.3: flowchart of the complete procedure

and is more suitable for a discrete setting.

The complete procedure of our method is illustrated in Figure 3.3 and given in Algorithm 4, where n_t is a realization of Gaussian noise N_t (or Binomial noise) when adopting the Gaussian (or Binomial) mechanism. D_H in Figure 3.3 represents the history data that can be used for estimating parameters but won't be revealed during this time horizon.

Note that the Gaussian/Binomial mechanism only specifies privacy parameters over one time step. In the next section we specify (ϵ_T, δ_T) over T steps.

3.5 Privacy Analysis

Next, we bound the total privacy loss when $X_{1:T}$ is released using Algorithm 4. Since the total privacy loss is accumulated during T steps, various composition methods can be applied to calculate (ϵ_T, δ_T). We use the moments accountant method from [4] when N_t is Gaussian; the corresponding result is given in Theorem 8. We use the composition theorem from [76] when N_t is Binomial with the corresponding result given in Theorem 9.

Theorem 8. *Let $Z_t = \mathcal{Q}(D_t)$ and Δ be the sensitivity of $\mathcal{Q}$, $\forall t$. Consider Algorithm 4 using zero-mean Gaussian noise with $Var(N_t) = \sigma_n^2$, $\forall t$, that takes sequence $z_{1:T}$ as input and outputs $x_{1:T}$. The following holds.*

(i) Given any $\epsilon_T \geq \frac{\Delta^2}{2\sigma_n^2}\sum_{t=1}^T w_t^2$*, the algorithm satisfies* (ϵ_T, δ_T)*-differential privacy for*

$$\delta_T = \exp\Big(\big(\frac{\frac{\Delta^2}{\sigma_n^2}\sum_{t=1}^T w_t^2}{4} - \frac{\epsilon_T}{2}\big)\big(\frac{\epsilon_T}{\frac{\Delta^2}{\sigma_n^2}\sum_{t=1}^T w_t^2} - \frac{1}{2}\big)\Big).$$

(ii) Given any $\delta_T \in (0,1)$*, the algorithm satisfies* (ϵ_T, δ_T)*-differential privacy for*

$$\epsilon_T = 2\sqrt{\frac{\Delta^2}{2\sigma_n^2}\sum_{t=1}^T w_t^2 \log(\frac{1}{\delta_T}) + \frac{\Delta^2}{2\sigma_n^2}\sum_{t=1}^T w_t^2}.$$

Theorem 8 says that if a sequence of noisy data is released following Algorithm 4 and the noise has variance σ_n^2, then with probability $1-\delta_T$, the total amount of privacy loss incurred to each individual during T time steps is bounded by ϵ_T. Here $\frac{\sigma_n}{\Delta}$ represents the degree of perturbation and w_t is the weight on the true value. Smaller perturbation and larger weight result in higher privacy loss. Because of the mapping between σ_n^2 and (ϵ_T, δ_T), we have the following result.

Corollary 1. *Let* $\{w_t\}_{t=1}^T$ *be the weights used in generating* $x_{1:T}$ *in Algorithm 4. To satisfy* (ϵ_T, δ_T)*-differential privacy, the variance of Gaussian noise should be:*

$$\sigma_n^2 \geq \frac{\Delta^2 \sum_{t=1}^T w_t^2}{2\epsilon_T + 4\ln\frac{1}{\delta_T} - 4\sqrt{(\ln\frac{1}{\delta_T})^2 + \epsilon_T \ln\frac{1}{\delta_T}}}.$$

To guarantee (ϵ_T, δ_T)-differential privacy, the noise magnitude will depend on both w_t and Δ. Larger sensitivity means larger impact of each individual on the released information and thus requires more perturbation for privacy protection; larger weights mean higher reliance on the true value in the released information, thus more perturbation is needed.

Theorem 9. *Let* $Z_t = \mathcal{Q}(D_t)$ *and* Δ *be the sensitivity of* $\mathcal{Q}$ $\forall t$*. Consider Algorithm 4 using Binomial noise with* $Var(N_t) = \frac{m}{2}$*,* $\forall t$ *that takes sequence* $z_{1:T}$ *as input and outputs* $x_{1:T}$*,* $\forall\ \tilde{\delta} \in [0,1]$*, if* $1 \leq w_t\Delta + \frac{2m+1}{\exp(\frac{\epsilon}{w_t\Delta})+1} \leq m+1$*,* $\forall t$*, then the algorithm is* $(\tilde{\epsilon}_{\tilde{\delta}}, 1-(1-\tilde{\delta})\prod_{t=1}^T(1-\delta_t))$*-differentially private for:*

$$\tilde{\epsilon}_{\tilde{\delta}} = \min\Big\{\sum_{t=1}^T \epsilon_t, \sum_{t=1}^T \frac{(e^{\epsilon_t}-1)\epsilon_t}{e^{\epsilon_t}+1} + \sqrt{\sum_{t=1}^T 2\epsilon_t^2 \log(e + \frac{\sqrt{\sum_{t=1}^T \epsilon_t^2}}{\tilde{\delta}})}, \sum_{t=1}^T \frac{(e^{\epsilon_t}-1)\epsilon_t}{e^{\epsilon_t}+1} + \sqrt{\sum_{t=1}^T 2\epsilon_t^2 \log(\frac{1}{\tilde{\delta}})}\Big\}$$

with any $\epsilon_t > 0$ *and corresponding*

$$\delta_t = \exp(-\frac{1}{m}(m - w_t\Delta + 1 - \frac{2m+1}{\exp(\frac{\epsilon_t}{w_t\Delta}) + 1})^2),$$

or with any $\delta_t \in (0,1)$ *and corresponding*

$$\epsilon_t = w_t\Delta \log(\frac{2m+1}{m - w_t\Delta + 1 - \sqrt{m \log \frac{1}{\delta_t}}} - 1).$$

Note that Algorithm 4 reduces to the baseline approach when $w_t = 1,\ \forall t$. Theorems 8, 9 and Corollary 1 also hold for the baseline method if we set $w_t = 1,\ \forall t$. When the noise variance is finite, using the baseline method we have $\forall \delta_T$, $\epsilon_T \to \infty$ as $T \to \infty$. However, under the proposed method, it is possible that $\lim_{T\to\infty} \epsilon_T < \infty$ by controlling w_t, e.g., by taking w_t as a decreasing geometric sequence.

3.6 Accuracy Analysis

In this section, we compare the accuracy of our method and the baseline method using the Mean Squared Error (MSE) measure, defined as $\mathbb{E}_{X_{1:T}}(\|x_{1:T} - z_{1:T}\|^2)$.

For simplicity of exposition, the analysis in this section is based on the assumption that the true values of parameters (ρ, μ, σ^2) of the underlying process are known. Additional error introduced by estimating parameters in Algorithm 3 is examined numerically in Section 3.8. In addition, we will only present the case of Gaussian AR(1) process and $\mathrm{Var}(N_t) = \sigma_n^2, \forall t$.

Theorem 10. *Let the sequence* $z_{1:T}$ *be generated by the Gaussian AR(1) process* $Z_{1:T}$ *with* $Z_t \sim \mathcal{N}(\mu, \sigma_z^2)$ *and* $Corr(Z_t Z_{t-\tau}) = \rho^{|\tau|},\ \forall t$. *Let* $x_{1:T}$ *be the sequence released by Algorithm 4. Then* $\mathbb{E}_{X_{1:T}}(\|x_{1:T} - z_{1:T}\|^2)$ *is given by*

$$\underbrace{\sigma_z^2(1 - \rho^2 \frac{\sigma_z^2}{\sigma_z^2 + \sigma_n^2}) \sum_{t=1}^{T} (1 - w_t)^2}_{\text{estimation error}} + \underbrace{T\sigma_n^2}_{\text{perturbation error}}.$$

Theorem 10 suggests that the total error consists of two parts: (i) estimation error and (ii) perturbation error. For the former, a sequence with stronger autocorrelation (larger ρ) enables more

accurate estimate, resulting in lower estimation error. Further, higher weight on the true value z_t (larger w_t), or less perturbation (smaller σ_n^2), also lowers the estimation error.

Theorem 11 below further compares the privacy-accuracy tradeoff of the two methods, where MSE is compared under the same privacy parameters (ϵ_T, δ_T).

Theorem 11. *Let sequential data $z_{1:T}$ be generated by the Gaussian AR(1) process $Z_{1:T}$ with $Z_t \sim \mathcal{N}(\mu, \sigma_z^2)$ and $Corr(Z_t Z_{t-\tau}) = \rho^{|\tau|}$, $\forall t$. Let $x_{1:T}^A$, $x_{1:T}^B$ be the sequences released by Algorithm 4 and the baseline method, respectively. Let $(\sigma_n^2)^A$, $(\sigma_n^2)^B$ be the corresponding noise variance. Suppose both outputs satisfy (ϵ_T, δ_T)-differential privacy, then*

$$\frac{T}{(\sigma_n^2)^B} = \frac{\sum_{t=1}^T w_t^2}{(\sigma_n^2)^A} = \frac{2\epsilon_T + 4\ln\frac{1}{\delta_T} - 4\sqrt{(\ln\frac{1}{\delta_T})^2 + \epsilon_T \ln\frac{1}{\delta_T}}}{\Delta^2} .$$

Furthermore, $\exists\, \{w_t\}_{t=1}^T, w_t \in (0,1)$ and $(\sigma_n^2)^A$, that satisfy Eqn. (3.3) and with which $x_{1:T}^A$ is more accurate than $x_{1:T}^B$.

Moreover, if a constant weight $w_t = w, \forall t$ is used, then $x_{1:T}^A$ is more accurate than $x_{1:T}^B$ if:

$$w > \frac{1 - (\sigma_n^2)^B/\sigma_z^2}{1 + (\sigma_n^2)^B/\sigma_z^2} . \tag{3.3}$$

As mentioned earlier, when $w_t = 1, \forall t$, Algorithm 4 reduces to the baseline method, and $x_{1:T}^A$ and $x_{1:T}^B$ become equivalent. Theorem 11 shows that our method can *strictly* improve the privacy-accuracy tradeoff by controlling $w_t \in (0,1)$. It also provides the guidance on how to select a constant weight $w_t = w, \forall t$, to guarantee this improvement from Eqn. (A.51): (i) If $(\sigma_n^2)^B > \sigma_z^2$, i.e., the privacy requirement is high and large perturbation is needed, then our method can always outperform the baseline regardless of the choice of $w \in (0,1)$. In particular, if choosing $w \to 0$, our method will have large estimation error, but privacy can be provided with insignificant perturbation; the overall error is dominated by the estimation error, which is still smaller than the perturbation error in the baseline. (ii) If $(\sigma_n^2)^B < \sigma_z^2$, then w should be sufficiently large to maintain accuracy.

3.7 Discussion

Generalization: The proposed method is not limited to AR(1) processes; it can be applied to any (weakly) stationary random process. This is because the LMMSE estimate only depends on the mean, variance and correlation of the random process. The methodologies used in Sections 3.5 and 3.6 are also not limited to AR(1) processes. In Section 3.8, the real-world datasets used in the experiments do not necessarily follow AR(1), but our method is shown to achieve better performance.

Robustness against certain attacks: Differential privacy is a strong privacy guarantee and a worst-case measure, as it bounds privacy loss over all possible outputs and inputs. In practice, how much information about $z_{1:T}$ can really be inferred by an attacker depends on how strong it is assumed to be. An attacker is able to infer more information with higher confidence if it knows the exact perturbation mechanism used in generating $x_{1:T}$, i.e., $\Pr(X_t|Z_t)$. If an attacker knows the noise distribution $\mathcal{N}(0,\sigma_n^2)$, then it will know $\Pr(X_t|Z_t)$ automatically with the baseline method, i.e., $X_t|Z_t \sim \mathcal{N}(Z_t,\sigma_n^2)$. However, with our method, $X_t|Z_t \sim \mathcal{N}(w_tZ_t+(1-w_t)\hat{Z}_t(x_{t-1}),\sigma_n^2)$. If w_t is *private*, thus unknown to the attacker, then $\Pr(X_t|Z_t)$ is not readily inferable. Therefore, in practice our method can prevent this class of attackers from knowing the details of the perturbation mechanism, thus can be stronger.

Impact of estimating parameters from noisy sequence: The analysis in Section 3.6 shows that when the true parameters of the underlying process are known, our algorithm can always outperform the baseline method. However, these may be unknown in reality and need to be estimated from the released sequence using Algorithm 3, which leads to additional estimation error. Nevertheless, this can still outperform the baseline method. Consider the extreme case where $(\sigma_n^2)^A \to +\infty$. The LMMSE estimate from the noisy data $\hat{Z}_t(x_{t-1}) \to \mathbb{E}(Z_t) \approx \hat{\mu}_{t-1}$. Since the added noise is zero-mean, with enough released data $\hat{\mu}_{t-1}$ can attain sufficient accuracy. Then x_t determined by both $\hat{\mu}_{t-1}$ and true z_t before adding noise becomes a filtered version of the true sequence, and its accuracy after adding noise will still be higher than the baseline method under the same privacy measure; this point is further validated by experiments in Section 3.8.

Other approaches to calibrating released sequence: We have used the convex combination of estimate $\hat{Z}_t(x_{t-1})$ and true data z_t to calibrate the released data. This method is effective and easy to use and analyze. In particular, the weight in the convex combination provides an additional degree of freedom and serves four purposes (Section 3.4). There are also other approaches to calibrating

the released sequence. For example, we can leverage all released points to estimate new data, and use a sequence of estimates to calibrate, i.e., $\sum_{i=1}^{t-1} w_i \hat{Z}_t(x_i) + w_t z_t$. One could also use a non-linear combination to calibrate, e.g., $w_t z_t + (1-w_t)\sqrt{z_t \hat{Z}_t(x_{t-1})}$.

3.8 Experiments

In this section, we compare the privacy-accuracy tradeoff of our method with other methods using real-world datasets. Fixed weights, $w_t = w$, $\forall t$, are used in the proposed method.

Methods: For comparison, in addition to the baseline method, we also consider the following.

- *Baseline-Laplace*: Laplace noise $n_t \sim Lap(0, \frac{T\Delta}{\epsilon_T})$ is added to z_t independently at each time step.
- *FAST without sampling* [41] [5]: Laplace noise $n_t \sim Lap(0, \frac{T\Delta}{\epsilon_T})$ is first added to z_t, then a posterior estimate of each z_t using the Kalman filter is released. Since it assumes the time series follows a random process $Z_{t+1} = Z_t + U_t$ with $U_t \sim \mathcal{N}(0, \sigma_u^2)$, to use the Kalman filter it requires σ_u^2 to be known in advance. Moreover, it also needs to use a Gaussian noise $\tilde{n}_t \sim \mathcal{N}(0, \sigma_{app}^2)$ to approximate the added Laplace noise n_t. In our experiments, σ_{app}^2 is chosen based on the guidelines provided in [41] and σ_u^2 that gives the best performance is selected using exhaustive search.
- *DFT* [121]: Discrete Fourier Transform is applied to the entire sequence first, then among T Fourier coefficients $DFT(z_{1:T})_j = \sum_{i=1}^{T} \exp(\frac{2\pi\sqrt{-1}}{T} ji) x_i$, $j \in [T]$, it selects the top d and perturbs each of them using Laplace noise $\frac{\sqrt{dT}\Delta}{\epsilon_T}$. Lastly, it pads $T-d$ 0's to this perturbed coefficients vector and applies Inverse Discrete Fourier Transform. In our experiments, d that gives the best performance is selected from $\{1, \cdots, T\}$ using exhaustive search.
- *BA* and *BD* [81]: Two privacy budget allocation mechanisms, Budget Distribution (BD) & Budget Absorption (BA), are used to dynamically allocate privacy budget over time based on the dissimilarity between the previously released data and the new data. The new private data is released at each time step only when the data is sufficiently different from the previously

[5] FAST samples $k < T$ points and allocates privacy budget ϵ_T to the sampled points. It adds Laplace noise $Lap(0, \frac{K\Delta}{\epsilon_T})$ to each sampled point and outputs the corresponding a posterior estimate, while for non-sampled points it outputs prior estimates. A similar sampling procedure can be added to our proposed method where we set $w_t = 0$ for non-sampled points.

released data; otherwise, the previous data is recycled and released again. The idea is to improve accuracy by allocating more privacy budgets to the most important data points.

Real-World Dataset: We use the following datesets in our experiments.

- *Ride-sharing counts* [42]: this is generated using historical log from Capital Bikeshare system, USA, in 2011. It includes the counts of rented bikes aggregated on both an hourly and daily basis.
- *NY traffic volume counts in 2011* [33]: this is collected by the Department of Transportation (DOT). It contains the counts of traffic in various roadways from 12AM to 1PM on an hourly basis each day. We aggregate the counts from all roadways and concatenate sequences from different days in chronological order.
- *Federal Test Procedure (FTP) drive cycle* [40]: this dataset includes a speed profile for vehicles, and it simulates urban driving patterns. It can be used for emission certification and fuel economy testing of vehicles in the United States.

Accuracy Metric: We use *relative error (RE)* defined as the normalized MSE to measure the accuracy of $x_{1:T}$:

$$RE(z_{1:T}, x_{1:T}) = \frac{1}{T} \frac{\|z_{1:T} - x_{1:T}\|_2}{\max_{1 \le t \le T} |z_t|} .$$

The comparison results are shown in Figure 3.4, where we use $\delta_T = 10^{-7}$ in baseline-Normal and the proposed method, and $\Delta = 1$ as each data point z_t is a count over a dataset. The left plot compares the relative error achieved by different methods under the same ϵ_T.

However, the baseline-Laplace, FAST and DFT methods satisfy $(\epsilon_T, 0)$-differential privacy while the baseline-Normal and proposed methods satisfy $(\epsilon_T, 10^{-7})$-differential privacy. Even though $\delta_T = 10^{-7}$ appears small, the total privacy loss ϵ_T under these methods are calculated using different composition methods. Comparing different methods solely based on ϵ_T may not be appropriate as the improvement in ϵ_T may come from the composition strategy but not the algorithm itself.

To address this issue, we add the right plot in Figure 3.4, where noises in baseline-Laplace and baseline-Normal are chosen such that the error achieved by baseline-Normal is no less than baseline-Laplace, i.e., the black curve is slightly over the green curve in the plot. This would guarantee that baseline-Normal provide stronger privacy than baseline-Laplace. By further controlling the proposed method to have the same privacy as baseline-Normal (noise variances in two methods satisfy Eqn. (3.3)), and FAST and DFT to have the same privacy as baseline-Laplace, we can

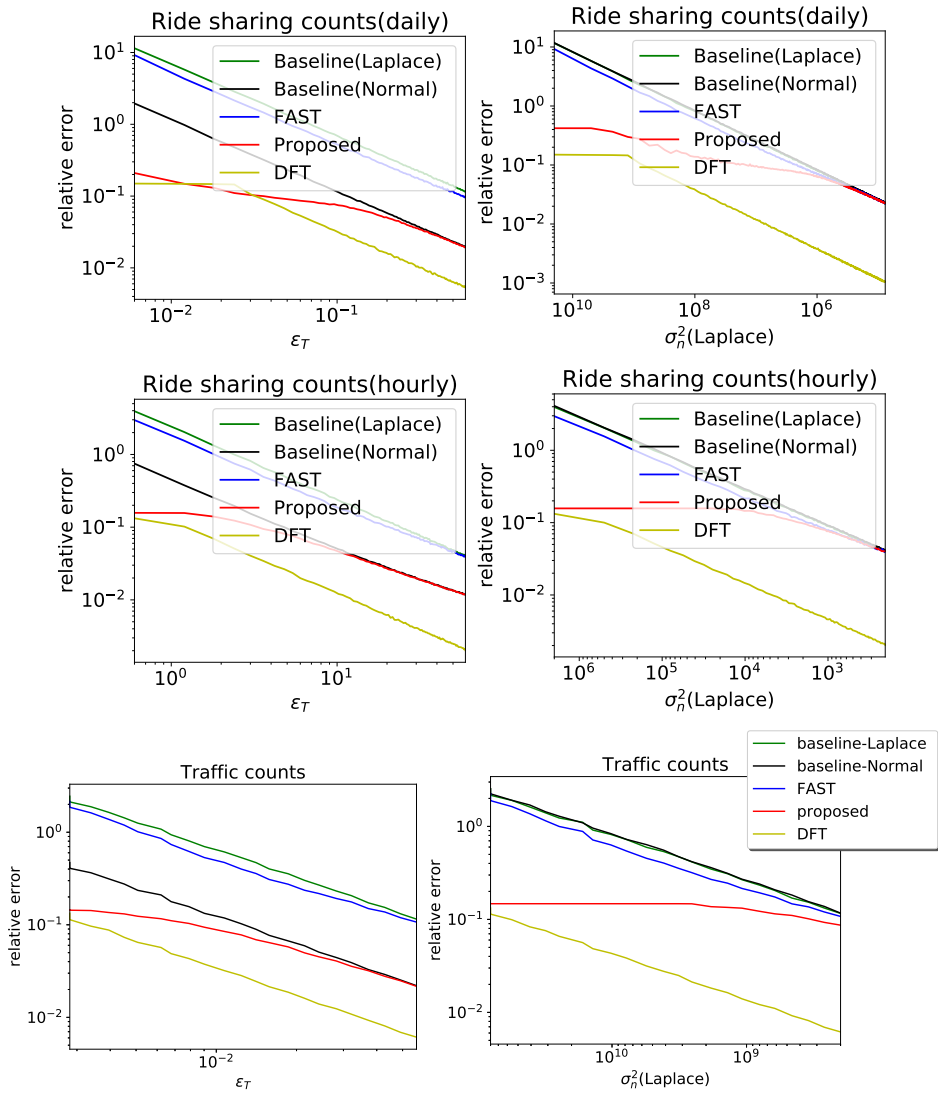


Figure 3.4: Comparison of different methods

guarantee the proposed method is at least as private as FAST and DFT. In the plot, the x-axis denotes the variance of added noise in baseline-Laplace and the noise parameters of the other methods are selected accordingly. It shows that the proposed method outperforms FAST; the improvement is more significant when the privacy requirement is high. While generally DFT performs better than the proposed method, it is an offline method which requires the entire sequence to be known a priori. However, as perturbation increases (more private), the proposed method can achieve similar

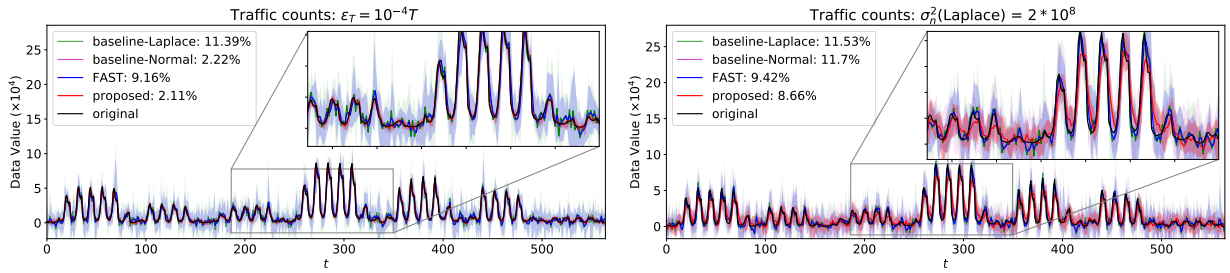


Figure 3.7: Sequences aggregated from 10 runs of experiments using different methods under the same ϵ_T (left plot). In the right plot, noise variance is selected in each method such that the proposed method and baseline-Normal are at least as private as FAST and baseline-Laplace.

performance as DFT.

The DFT method can also be adapted online; one way to do this is to perform DFT over a subsequence of length $T_{delay} \ll T$ (data released with delay T_{delay}). We examine the performance of such a method on the Traffic dataset by comparing it with DFT and baseline-Laplace. Figure 3.5 shows that when $T_{delay} = 0$ (data released in real-time, DFT applied to one data point each time and on one coefficient), the performance is similar to baseline-Laplace; as T_{delay} increases, its accuracy increases at the expense of increased delay.

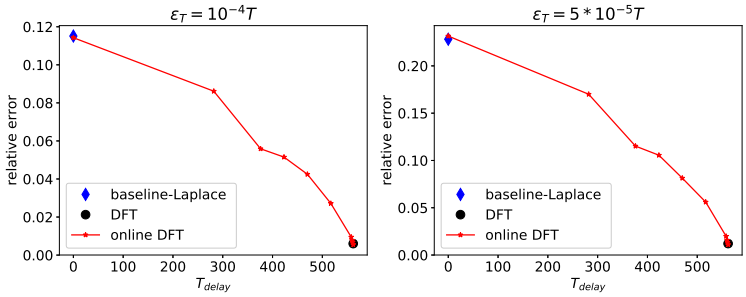


Figure 3.5: Comparison with Online DFT

Figure 3.7 shows the private traffic counts generated using various methods. For each method, we repeat the experiment 10 times and obtain 10 sample paths $\{x_{1:T}\}_{k=1}^{10}$. The curves in the plot show the average $\frac{1}{10}\sum_{k=1}^{10} x_{1:T}^k$ while the shaded area indicates their variance whose upper and lower bound at each t are $\max_k x_t^k$ and $\min_k x_t^k$, respectively. The similar results can be observed for other datasets.

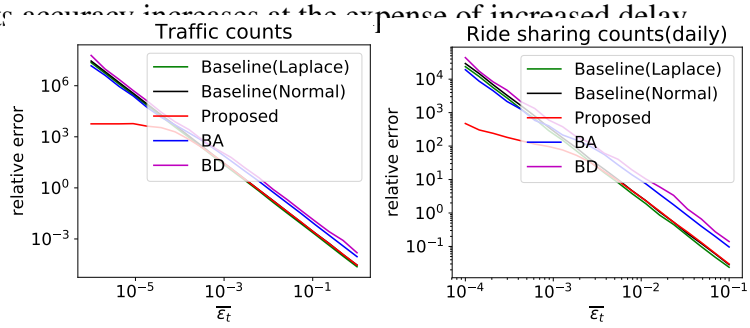


Figure 3.6: Comparison with BA and BD [81].

We also compare our proposed method with BA and BD proposed in [81]. Unlike our model,

where a single query is released at every time step, BA and BD are designed to release a vector of length d each time. Moreover, BA and BD adopt $(\epsilon, 0)$-differential privacy. In order to compare with our method, we set $d = 1$ and use baseline-Laplace and baseline-Normal as two baselines. Specifically, we choose noises for different methods such that: (1) our proposed method and baseline-Normal have the same privacy guarantee; (2) BA, BD, and baseline-Laplace have the same privacy guarantee; and (3) baseline-Normal is at least as private as baseline-Laplace. The results are shown in Figure 3.6, where the y-axis indicates the averaged relative error of 10 independent runs of experiment and x-axis is the privacy loss per time step under baseline-Laplace. As illustrated, our method outperforms others. It is worth noting that BA and BD may not even outperform baseline-Laplace. This is because in both BA and BD, half of the privacy budget is assigned to measure the dissimilarity between previously released data and new data; thus only half of the privacy budget is left for releasing the sequence. Moreover, as mentioned, BA and BD are meant for releasing a vector, especially when d is large; the error of the released sequence can be large when d is small (Theorem 6 and 7 in [81]). It further suggests that in settings where only a single query is released ($d = 1$), BA and BD may not be suitable.

As mentioned earlier, the baseline is a special case ($w_t = 1$, $\forall t$) of our method, which can always outperform the former with better tuned weights. The achievable improvement depends on the correlation of the sequence. We show this in Figure 3.8, the error of various synthetic sequences using different weights under the same privacy ϵ_T. Each sequence follows Gaussian AR(1) with $Z_t \sim \mathcal{N}(0, 1)$ but the correlation ρ varies from 0.1 to 0.9. It shows that (i) in all cases, one can find weights for our method to outperform the baseline; sequences with high ρ have the highest accuracy under the same ϵ_T; (ii) with weak (resp. strong) privacy as shown on the right (resp. left), the smallest weights that can give improvement are close to 1 (resp. 0) and the achievable improvement is small (resp. large) as compared to the baseline. As released data depends less (resp. more) on estimates when weights are large (resp. small), the correlation within the sequence does not (resp. does) affect performance significantly. In the right (resp. left) plot with weak (resp. strong) privacy, curves with lowest error are similar under different ρ (resp. decreases in ρ).

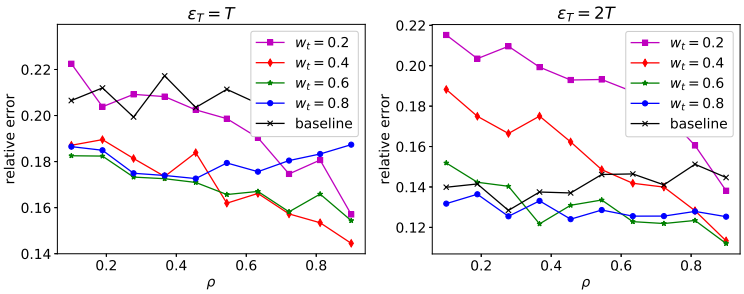


Figure 3.8: Impact of correlation on performance

We also examine the impact of estimating parameters from a noisy sequence; the result is shown in Figure 3.9, where Gaussian AR(1) sequences are generated. Red curves represent the relative error achieved using the proposed method where $\hat{\mu}_t$, $\hat{\sigma}_t^2$ and $\hat{\rho}_t$ at each time are estimated from the previous released sequence; blue curves represent the case where we use true parameters μ, σ^2, ρ to estimate z_t using x_{t-1}. As expected, estimating parameters from a noisy sequence degrades the performance. However, even with this impact the proposed method continues to outperform the baseline significantly.

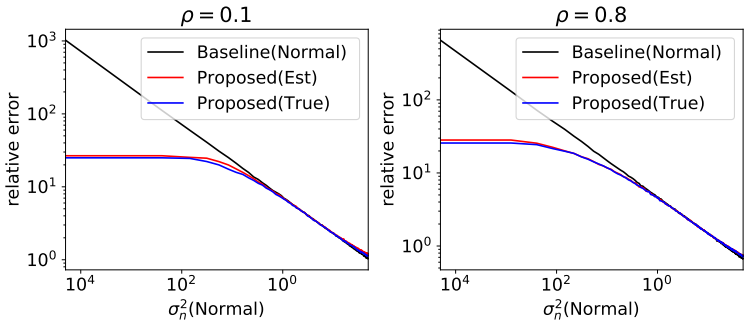


Figure 3.9: Impact of estimation from noisy sequence: $Z_{1:T}$ satisfies $Z_{t+1} = \rho Z_t + U_t$ with $U_t \sim \mathcal{N}(0, 10)$, $Z_0 = 0$ and weak ($\rho = 0.1$) or strong ($\rho = 0.8$) autocorrelation.

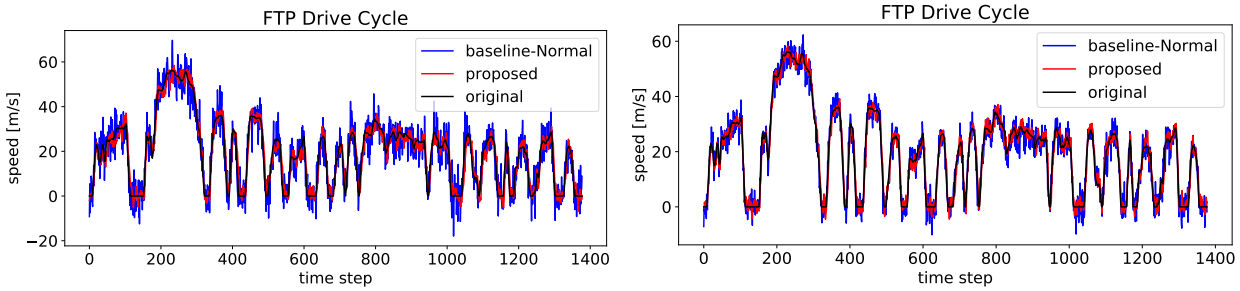


Figure 3.10: Drive cycles under different levels of privacy: the privacy guarantee in the left plot is stronger than that of the right plot.

The proposed method is not limited to count queries and is more broadly applicable. For example, this method can be used in intelligent transportation systems to enable *private* vehicle-to-vehicle communication. In our studies [67, 148], a predictive cruise controller is designed for a follower vehicle using a private speed profile transmitted from its leader vehicle. Specifically, instead of broadcasting the real speed profile (FTP drive cycle), the leader vehicle generates a differentially private speed profile using the proposed method. A follower vehicle then designs an optimal speed planner based on the received information. Within this application context, query $\mathcal{Q}(v)$ represents the vehicle's speed information, and sensitivity Δ is the range of the vehicle's speed. Figure 3.10

shows the private speed profiles generated using the proposed method and baseline-Normal. A private optimal speed planner is designed in [67, 148] using these private profiles. The results show that the controller performance deteriorates significantly under the baseline method. In contrast, the controller designed with the proposed method can attain an accuracy that is sufficient for predictive control purposes. We refer an interested reader to [67, 148] for more details and the performance of the private controller.

Part II

Fair Machine Learning with Human in Feedback Loops

CHAPTER 4

Long-Term Impact of Fairness Interventions on Group Representation

4.1 Introduction

From this chapter, we set out to understand the long-term impact of (fair) machine learning algorithms on the well-being of different social groups. We will begin with the dynamics where individuals interact with ML decisions by leaving/staying ML systems. Specifically, we consider a discrete-time sequential decision process applied to a population consists of individuals from multiple demographic groups, where individuals' responses (stay or leave) to the decisions made at each time step are manifested in changes in the population in the next time step.

To motivate our problem, consider an example in speech recognition. It has been documented that speech recognition products such as Amazon's Alexa and Google Home have accent bias against non-native speakers [59], with native speakers experience much higher quality than non-native speakers. If this difference leads to more native speakers using such products while driving away non-native speakers, then over time the data used to train the model may become even more skewed toward native speakers, with fewer and fewer non-native samples. Without intervention, the resulting model becomes even more accurate for the former and less for the latter, which then reinforces their respective user experience [60].

In this chapter, we are particularly interested in understanding what happens to group representation (demographic sizes) over time when models with fairness guarantee (e.g., `EqOpt`, `DP`) are used, and how it is affected when the groups' underlying feature distributions are also affected/reshaped by decisions.

Our main contributions and findings are as follows.

1. We introduce a user retention model to capture users' reaction (stay or leave) to the decision (Section 4.3).
2. We first consider the case where the decisions only affect individual's participation and the groups' feature distributions are fixed over time.
    - We identify conditions under which group representation disparity exacerbates over time and eventually the disadvantaged groups may diminish entirely from the system (Theorems 12 and 13).
    - We show the conditions that lead to exacerbation of disparity unfortunately can be easily satisfied when decisions are made based on a typical algorithm (e.g., taking objective as minimizing the total loss) under commonly used fairness criteria (e.g., `DP`, `EpOpt`) (Theorem 15).
3. We further consider the case where the decisions also affect groups' feature distributions, and examine its impact on group representations.
    - We show that exacerbation of disparity continues to hold and can *accelerate* when feature distributions are affected and change over time (Theorem 16).
4. Our results show that if the factors equalized by the fairness criterion do not match what drives user retention, then the difference in (perceived) treatment will exacerbate representation disparity over time. It suggests fairness has to be defined with a good understanding of how users are affected by the decisions.
5. Given the knowledge of user dynamics, we propose a method for finding the proper fairness criterion that mitigates representation disparity (Section 4.4.4).

The remainder of this chapter is organized as follows. Section 4.2 presents related work. Section 4.3 formulates the problem. The impact of various fairness criteria on group representation disparity is analyzed and presented in Section 4.4, as well as potential mitigation. Experiments are presented in Section 4.5. All proofs can be found in Appendix C.

4.2 Related Work

The impact of fairness interventions on both individuals and society, and the fairness in sequential decision making have been studied extensively in the literature [60,63,65,79,103]. Specifically, [103] constructs a one-step feedback model over two consecutive time steps and characterizes the impact

of fairness criteria (statistical parity and equal of opportunity) on changing each individual's feature and reshaping the entire population. Similarly, [63] proposes an effort-based measure of unfairness and constructs an individual-level model characterizing how an individual responds to the decisions based on it. The impact on the entire group is then derived from it and the impacts of fairness intervention are examined. While both highlight the importance of temporal modeling in evaluating the fairness, their main focus is on the adverse impact on feature distribution, rather than on group representation disparity. In contrast, our work focuses on the latter but also considers the impact of reshaping feature distributions. Moreover, we formulate the long-term impact over infinite horizon while [63, 103] only inspect the impact over two steps.

[60] also considers a sequential framework where the user departure is driven by model accuracy. It adopts the objective of minimizing the loss of the group with the highest loss (instead of overall or average loss), which can prevent the extinction of any group from the system. It requires multiple demographic groups use the same model and does not adopt any fairness criterion. In contrast, we are more interested in the impact of various fairness criteria on representation disparity and if it is possible to sustain the group representation by imposing any fairness criterion. Other differences include the fact we consider the case when feature distributions are reshaped by the decisions (Section 4.4.3) and [60] does not.

[79] constructs a two-stage model in the context of college admission, it shows that increasing admission rate of a group can increase the overall qualification for this group overtime. [65] describes a model in the context of labor market. They show that imposing the demographic parity constraint can incentivize under-represented groups to invest in education, which leads to a better long-term equilibrium.

Another line of research is to study fairness problems in online learning [16, 31, 62, 70, 131, 144]. Most of them focus on proposing appropriate fairness notions to improve the fairness-accuracy trade-off. To the best of our knowledge, none of them considers the impact of fairness criteria on group representation disparity.

4.3 Problem Formulation

Consider two demographic groups $\mathcal{G}_a$, $\mathcal{G}_b$ distinguished based on some sensitive attribute $S \in \{a, b\}$ (e.g., gender, race). An individual from either group has feature $X_t \in \mathbb{R}^d$ and label $Y_t \in \{0, 1\}$ at time t, both can be time varying. Denote by $\mathcal{G}_s^y \subset \mathcal{G}_s$ the subgroup with label y, $y \in \{0, 1\}$, $s \in \{a, b\}$, $f_{s,t}^y(x) :=$

$P_{X_t|Y_t,S}(x|y,s)$ its feature distribution and $n_s(t)$ the size of $\mathcal{G}_s$ as a fraction of the entire population at time t. Then the difference between $n_a(t)$ and $n_b(t)$ measures the representation disparity between two groups at time step t. Denote by $\alpha_{s,t} = P_{Y_t|S}(1|s)$ the fraction of positive label in group s at time t, then the distribution of X over $\mathcal{G}_s$ is given by $f_{s,t}(x) = P_{X_t|S}(x|s) = \alpha_{s,t} f^1_{s,t}(x) + (1-\alpha_{s,t}) f^0_{s,t}(x)$ and $f_{a,t} \neq f_{b,t}$.

Consider a sequential setting where the decision maker at each time makes a decision $D_t \in \{0,1\}$ about each individual based on group-dependent policies. Let $\pi_{s,t}(x) = P_{D_t|X_t,S}(1|x_t,s)$ be the decision rule for $\mathcal{G}_s$ at time t, which is parameterized by $\theta_s(t) \in \mathbb{R}^d$, $s \in \{a,b\}$. The goal of the decision maker at time t is to find the best parameters $\theta_a(t)$, $\theta_b(t)$ such that the corresponding decisions about individuals from $\mathcal{G}_a$, $\mathcal{G}_b$ maximize its utility (or minimize its loss) in the current time. Within this context, the commonly studied fair machine learning problem is the one-shot problem stated as follows, at time step t:

$$\begin{aligned} \min_{\theta_a,\theta_b} \quad & \boldsymbol{O}_t(\theta_a,\theta_b;n_a(t),n_b(t)) = n_a(t)O_{a,t}(\theta_a) + n_b(t)O_{b,t}(\theta_b) \\ \text{s.t.} \quad & \Gamma_{\mathcal{C},t}(\theta_a,\theta_b) = 0\ , \end{aligned} \tag{4.1}$$

where $\boldsymbol{O}_t(\theta_a,\theta_b;n_a(t),n_b(t))$ is the overall objective of the decision maker at time t, which consists of sub-objectives from two groups weighted by their group proportions.[1] $\Gamma_{\mathcal{C},t}(\theta_a,\theta_b) = 0$ characterizes fairness constraint $\mathcal{C}$, which requires the parity of certain statistical measure (e.g., positive classification rate, false positive rate, etc.) across different demographic groups. Some commonly used criteria will be elaborated in Section 4.4.1. Both $O_{s,t}(\theta_s)$ and $\Gamma_{\mathcal{C},t}(\theta_a,\theta_b) = 0$ depend on $f_{s,t}(x)$. The resulting solution $(\theta_a(t),\theta_b(t))$ will be referred to as the one-shot fair decision under fairness $\mathcal{C}$, where the optimality only holds for a single time step t.

In this study, we seek to understand how the group representation evolves in a sequential setting over the long run when different fairness criteria are imposed. To do so, the impact of the current decision on the size of the underlying population is modeled by the following discrete-time retention dynamics. Denote by $N_s(t) \in \mathbb{R}_+$ the expected number of users in group s at time t:

$$N_s(t+1) = N_s(t) \cdot \lambda_{s,t}(\theta_s(t)) + \beta_s\ , \forall s \in \{a,b\}, \tag{4.2}$$

[1] This is a typical formulation if the objective $\boldsymbol{O}_t$ measures the average performance of decisions over all samples, i.e., $\boldsymbol{O}_t = \frac{1}{|\mathcal{G}_a|+|\mathcal{G}_b|}(\sum_{i\in\mathcal{G}_a} O^i_t + \sum_{i\in\mathcal{G}_b} O^i_t) = \frac{1}{|\mathcal{G}_a|+|\mathcal{G}_b|}(|\mathcal{G}_a|O_{a,t} + |\mathcal{G}_b|O_{b,t})$, where O^i_t measures the performance of each sample i and $O_{k,t} = \frac{1}{|\mathcal{G}_k|}\sum_{i\in\mathcal{G}_k} O^i_t$ is the average performance of $\mathcal{G}_k$.

where $\lambda_{s,t}(\theta_s(t))$ is the retention rate, i.e., the probability of a user from $\mathcal{G}_s$ who was in the system at time t remaining in the system at time $t+1$. This is assumed to be a function of the user experience, which could be the actual accuracy of the algorithm or their perceived (mis)treatment. This experience is determined by the application and is different under different contexts. For instance, in domains of speaker verification and medical diagnosis, it can be considered as the average loss, i.e., a user stays if he/she can be classified correctly; in loan/job application scenarios, it can be the rejection rates, i.e., user stays if he/she gets approval. β_s is the expected number of exogenous arrivals to $\mathcal{G}_s$ and is treated as a constant in our analysis, though our main conclusion holds when this is modeled as a random variable. Accordingly, the relative group representation for time step $t+1$ is updated as

$$n_s(t+1) = \frac{N_s(t+1)}{N_a(t+1) + N_b(t+1)}, \forall s \in \{a, b\}.$$

For the remainder of this chapter, $\frac{n_a(t)}{n_b(t)}$ is used to measure the group representation disparity at time t. As $n_s(t)$ and $f_{s,t}(x)$ change over time, the one-shot problem (4.1) is also time varying. In the next section, we examine what happens to $\frac{n_a(t)}{n_b(t)}$ when one-shot fair decisions are applied in each step.

4.4 Group Representation Disparity in the Sequential Setting

Below we present results on the monotonic change of $\frac{n_a(t)}{n_b(t)}$ when applying one-shot fair decisions in each step. It shows that the group representation disparity can worsen over time and may lead to the extinction of one group under a monotonicity condition stated as follows.

Monotonicity Condition. *Consider two one-shot problems defined in* (4.1) *with objectives* $\widehat{\boldsymbol{O}}(\theta_a, \theta_b; \widehat{n}_a, \widehat{n}_b)$ *and* $\widetilde{\boldsymbol{O}}(\theta_a, \theta_b; \widetilde{n}_a, \widetilde{n}_b)$ *over distributions* $\widehat{f_k}(x)$, $\widetilde{f_k}(x)$ *respectively. Let* $(\widehat{\theta}_a, \widehat{\theta}_b)$, $(\widetilde{\theta}_a, \widetilde{\theta}_b)$ *be the corresponding fair decisions. We say that two problems* $\widehat{\boldsymbol{O}}$ *and* $\widetilde{\boldsymbol{O}}$ *satisfy the monotonicity condition given a dynamic model if for any* $\widehat{n}_a + \widehat{n}_b = 1$ *and* $\widetilde{n}_a + \widetilde{n}_b = 1$ *such that* $\frac{\widehat{n}_a}{\widehat{n}_b} < \frac{\widetilde{n}_a}{\widetilde{n}_b}$, *the resulting retention rates satisfy* $\widehat{\lambda}_a(\widehat{\theta}_a) < \widetilde{\lambda}_a(\widetilde{\theta}_a)$ *and* $\widehat{\lambda}_b(\widehat{\theta}_b) > \widetilde{\lambda}_b(\widetilde{\theta}_b)$.

Note that this condition is defined over two one-shot problems and a given dynamic model. It is not limited to specific families of objective or constraint functions; nor is it limited to one-dimensional features. The only thing that matters is the group proportions within the system and the retention rates determined by the decisions and the dynamics. It characterizes a situation where

when one group's representation increases, the decision becomes more in favor of this group and less favorable to the other, so that the retention rate is higher for the favored group and lower for the other.

Theorem 12. *[Exacerbation of representation disparity] Consider a sequence of one-shot problems* (4.1) *with objective* $\boldsymbol{O}_t(\theta_a, \theta_b; n_a(t), n_b(t))$ *at each time t. Let* $(\theta_a(t), \theta_b(t))$ *be the corresponding solution and* $\lambda_{s,t}(\theta_s(t))$ *be the resulting retention rate of* $\mathcal{G}_s$*,* $s \in \{a, b\}$ *under a dynamic model* (4.2)*. If the initial states satisfy* $\frac{N_a(1)}{N_b(1)} = \frac{\beta_a}{\beta_b}$*,* $N_s(2) > N_s(1)$*,*[2] *and one-shot problems in any two consecutive time steps, i.e.,* $\boldsymbol{O}_t$*,* $\boldsymbol{O}_{t+1}$*, satisfy monotonicity condition under the given dynamic model, then the following holds. Let* $\diamond$ *denote either " < " or " = " or " > ", if* $\lambda_{a,1}(\theta_a(1)) \diamond \lambda_{b,1}(\theta_b(1))$*, then* $\frac{n_a(t+1)}{n_b(t+1)} \diamond \frac{n_a(t)}{n_b(t)}$ *and* $\lambda_{a,t+1}(\theta_a(t+1)) \diamond \lambda_{a,t}(\theta_a(t)) \diamond \lambda_{b,t}(\theta_b(t)) \diamond \lambda_{b,t+1}(\theta_b(t+1))$*,* $\forall t$*.*

Theorem 12 says that once a group's proportion starts to change (increase or decrease), it will continue to change in the same direction. This is because under the monotonicity condition, there is a feedback loop between representation disparity and the one-shot decisions: the former drives the latter which results in different user retention rates in the two groups, which then drives future representation.

The monotonicity condition can be satisfied under some commonly used objectives, dynamics and fairness criteria. This is characterized in the following theorem.

Theorem 13. *[A case satisfying monotonicity condition] Consider two one-shot problems defined in* (4.1) *with objectives* $\widetilde{\mathbf{O}}(\theta_a, \theta_b; \widehat{n}_a, \widehat{n}_b) = \widehat{n}_a O_a(\theta_a) + \widehat{n}_b O_b(\theta_b)$ *and* $\widehat{\mathbf{O}}(\theta_a, \theta_b; \widetilde{n}_a, \widetilde{n}_b) = \widetilde{n}_a O_a(\theta_a) + \widetilde{n}_b O_b(\theta_b)$ *over the same distribution* $f_s(x)$ *with* $\widehat{n}_a + \widehat{n}_b = 1$ *and* $\widetilde{n}_a + \widetilde{n}_b = 1$*. Let* $(\widehat{\theta}_a, \widehat{\theta}_b)$*,* $(\widetilde{\theta}_a, \widetilde{\theta}_b)$ *be the corresponding solutions. Under the condition that* $O_s(\widehat{\theta}_s) \neq O_s(\widetilde{\theta}_s)$ *for all possible* $\widehat{n}_s \neq \widetilde{n}_s$*, if the dynamics satisfy* $\lambda_s(\theta_s) = h_s(O_s(\theta_s))$ *for some decreasing function* $h_s(\cdot)$*, then* $\widetilde{\mathbf{O}}$ *and* $\widehat{\mathbf{O}}$ *satisfy the monotonicity condition.*

The above theorem identifies a class of cases satisfying the monotonicity condition; these are cases where whenever the group proportion changes, the decision will cause the sub-objective function value to change as well, and the sub-objective function value drives user departure.

For the rest of the chapter we will focus on the one-dimensional setting. Some of the cases we consider are special cases of Theorem 13 (Section 4.4.2). Others such as the time-varying feature distribution $f_{s,t}(x)$ considered in Section 4.4.3 also satisfy the monotonicity condition but are not captured by Theorem 13.

[2]This condition will always be satisfied when the system starts from a near empty state.

4.4.1 The One-Shot Problem

Consider a binary classification problem based on feature $X \in \mathbb{R}$. Let decision rule $\pi_s(x) = \mathbf{1}(x \geq \theta_s)$ be a threshold policy parameterized by $\theta_s \in \mathbb{R}$ and $L(y, \pi_s(x)) = \mathbf{1}(y \neq \pi_s(x))$ the 0-1 loss incurred by applying threshold θ on individuals with data (x, y).

The goal of the decision maker at each time is to find a pair $(\theta_a(t), \theta_b(t))$ subject to criterion $\mathcal{C}$ such that the total expected loss is minimized, i.e., $\boldsymbol{O}_t(\theta_a, \theta_b; n_a(t), n_b(t)) = n_a(t)L_{a,t}(\theta_a) + n_b(t)L_{b,t}(\theta_b)$, where $L_{s,t}(\theta_s) = \alpha_{s,t} \int_{-\infty}^{\theta_s} f_{s,t}^1(x)dx + (1 - \alpha_{s,t}) \int_{\theta_s}^{\infty} f_{s,t}^0(x)dx$ is the expected loss $\mathcal{G}_s$ experiences at time t. Some examples of $\Gamma_{\mathcal{C},t}(\theta_a, \theta_b)$ are as follows.

1. Simple fair (`Simple`): $\Gamma_{\texttt{Simple},t} = \theta_a - \theta_b$. Imposing this criterion simply means we ensure the same decision parameter is used for both groups.
2. Equal opportunity (`EqOpt`): $\Gamma_{\texttt{EqOpt},t} = \int_{\theta_a}^{\infty} f_{a,t}^0(x)dx - \int_{\theta_b}^{\infty} f_{b,t}^0(x)dx$. This requires the false positive rate (FPR) be the same for different groups,[3] i.e., $\Pr(\pi_a(X) = 1 | Y = 0, S = a) = \Pr(\pi_b(X) = 1 | Y = 0, S = b)$.
3. Demographic parity (`DP`): $\Gamma_{\texttt{DP},t} = \int_{\theta_a}^{\infty} f_{a,t}(x)dx - \int_{\theta_b}^{\infty} f_{b,t}(x)dx$. This requires different groups be given equal probability of being labelled 1, i.e., $\Pr(\pi_a(X) = 1 | S = a) = \Pr(\pi_b(X) = 1 | S = b)$.
4. Equalized loss (`EqLos`): $\Gamma_{\texttt{EqLos},t} = L_{a,t}(\theta_a) - L_{b,t}(\theta_b)$. This requires that the expected loss across different groups be equal.

Notice that for `Simple`, `EqOpt` and `DP` criteria, the following holds: $\forall t$, (θ_a, θ_b), and (θ'_a, θ'_b) that satisfy $\Gamma_{\mathcal{C},t}(\theta_a, \theta_b) = \Gamma_{\mathcal{C},t}(\theta'_a, \theta'_b) = 0$, we have $\theta_a \geq \theta'_a$ if and only if $\theta_b \geq \theta'_b$.

Some technical assumptions on the feature distributions are in order.

1. For $y \in \{0, 1\}$, $f_{a,t}^y(x), f_{b,t}^y(x)$ have bounded support on $[\underline{a}_t^y, \overline{a}_t^y]$ and $[\underline{b}_t^y, \overline{b}_t^y]$ respectively.
2. $f_{s,t}^1(x)$ and $f_{s,t}^0(x)$ overlap, i.e., $\underline{a}_t^0 < \underline{a}_t^1 < \overline{a}_t^0 < \overline{a}_t^1$ and $\underline{b}_t^0 < \underline{b}_t^1 < \overline{b}_t^0 < \overline{b}_t^1$.

The main technical assumption is stated as follows.

Assumption 7. *Let $\mathcal{T}_{s,t} = [\underline{s}_t^1, \overline{s}_t^0]$ be the overlapping interval between $f_{s,t}^0(x)$ and $f_{s,t}^1(x)$. Distribution $f_{s,t}^1(x)$ is strictly increasing and $f_{s,t}^0(x)$ is strictly decreasing over $\mathcal{T}_{s,t}$, $\forall s \in \{a, b\}$.*

For bell-shaped feature distributions (e.g., Normal, Cauchy, etc.), Assumption 7 implies that $f_{s,t}^1(x)$ and $f_{s,t}^0(x)$ are sufficiently separated. As we show later, this assumption helps us establish the monotonic convergence of thresholds $(\theta_a(t), \theta_b(t))$ but is not necessary for the convergence of group

[3] Depending on the context, this criterion can also refer to equal false negative rate (FNR), true positive rate (TPR), or true negative rate (TNR), but the analysis is essentially the same.

representation. We next find the one-shot decision to this problem under `Simple`, `EqOpt`, and `DP` fairness criteria.

Lemma 5. *Under Assumption 7, $\forall s \in \{a, b\}$, the optimal threshold at time t for $\mathcal{G}_k$ without constraint $\mathcal{C}$ is*

$$\theta_s^*(t) = \arg\min_{\theta_s} L_{s,t}(\theta_s) = \begin{cases} \underline{s}_t^1, \text{ if } \alpha_{s,t} f_{s,t}^1(\underline{s}_t^1) \geq (1-\alpha_{s,t}) f_{s,t}^0(\underline{s}_t^1) \\ \delta_{s,t}, \text{ if } \alpha_{s,t} f_{s,t}^1(\underline{s}_t^1) < (1-\alpha_{s,t}) f_{s,t}^0(\underline{s}_t^1) \ \& \ \alpha_{s,t} f_{s,t}^1(\overline{s}_t^0) > (1-\alpha_{s,t}) f_{s,t}^0(\overline{s}_t^0) \\ \overline{s}_t^0, \text{ if } \alpha_{s,t} f_{s,t}^1(\overline{s}_t^0) \leq (1-\alpha_{s,t}) f_{s,t}^0(\overline{s}_t^0) \end{cases}$$

where $\delta_{s,t} \in \mathcal{T}_{s,t}$ is defined such that $\alpha_{s,t} f_{s,t}^1(\delta_{s,t}) = (1-\alpha_{s,t}) f_{s,t}^0(\delta_{s,t})$. Moreover, $L_{s,t}(\theta_k)$ is decreasing in θ_s over $[\underline{s}_t^0, \theta_s^*(t)]$ and increasing over $[\theta_s^*(t), \overline{s}_t^1]$.

Below we will focus on the case when $\theta_a^*(t) = \delta_{a,t}$ and $\theta_b^*(t) = \delta_{b,t}$, while analysis for the other cases are essentially the same. For `Simple`, `DP` and `EqOpt` fairness, $\exists$ a strictly increasing function $\eta_t^{\mathcal{C}}$, such that $\Gamma_{\mathcal{C},t}(\eta_t^{\mathcal{C}}(\theta_b), \theta_b) = 0$. Denote by $(\eta_t^{\mathcal{C}})^{-1}$ the inverse of $\eta_t^{\mathcal{C}}$. Without loss of generality, we will assign group labels a and b such that $\eta_t^{\mathcal{C}}(\delta_{b,t}) < \delta_{a,t}$ and $(\eta_t^{\mathcal{C}})^{-1}(\delta_{a,t}) > \delta_{b,t}$, $\forall t$. [4]

Lemma 6. *Under `Simple`, `EqOpt`, `DP` fairness criteria, one-shot fair decision at time t satisfies $(\theta_a^*(t), \theta_b^*(t)) = \arg\min_{\theta_a, \theta_b} n_a(t) L_{a,t}(\theta_a) + n_b(t) L_{b,t}(\theta_b) \in \{(\theta_a, \theta_b) | \theta_a \in [\eta_t^{\mathcal{C}}(\delta_{b,t}), \delta_{a,t}], \theta_b \in [\delta_{b,t}, (\eta_t^{\mathcal{C}})^{-1}(\delta_{a,t})], \Gamma_{\mathcal{C},t}(\theta_a, \theta_b) = 0\} \neq \emptyset$ regardless of group proportions $n_a(t), n_b(t)$.*

Lemma 6 shows that given feature distributions $f_{a,t}(x)$, $f_{b,t}(x)$, although one-shot fair decisions can be different under different group proportions $n_a(t), n_b(t)$, these solutions are all bounded by the same compact intervals (Figure 4.1). Theorem 14 below describes the more specific relationship between group representation $\frac{n_a(t)}{n_b(t)}$ and the corresponding one-shot decision $(\theta_a(t), \theta_b(t))$.

Theorem 14. *[Impact of group representation disparity on the one-shot decision] Consider the one-shot problem with group proportions $n_a(t), n_b(t)$ at time step t, let $(\theta_a(t), \theta_b(t))$ be the corresponding one-shot decision under either `Simple`, `EqOpt` or `DP` criterion. Under Assumption 7, $(\theta_a(t), \theta_b(t))$ is unique and satisfies the following:*

$$\Psi_{\mathcal{C},t}(\theta_a(t), \theta_b(t)) = \frac{n_a(t)}{n_b(t)}, \tag{4.3}$$

[4] If the change of $f_{a,t}(x)$ and $f_{b,t}(x)$ w.r.t. the decisions follows the same rule (e.g., examples given in Section 4.4.3), then this relationship holds $\forall t$.

where $\Psi_{\mathcal{C},t}$ is some function increasing in $\theta_a(t)$ and $\theta_b(t)$, with details illustrated in Table 4.1.

	$\theta_a \in [\underline{a}_t^0, \underline{a}_t^1]$, $\theta_b \in \mathcal{T}_{b,t}$	$\theta_a \in \mathcal{T}_{a,t}$, $\theta_b \in \mathcal{T}_{b,t}$	$\theta_a \in \mathcal{T}_{a,t}$, $\theta_b \in [\overline{b}_t^0, \overline{b}_t^1]$
`EqOpt`	$\left(\frac{\alpha_{b,t}}{1-\alpha_{b,t}} \frac{f_{b,t}^1(\theta_b)}{f_{b,t}^0(\theta_b)} - 1\right) \frac{1-\alpha_{b,t}}{1-\alpha_{a,t}}$	$\frac{\frac{\alpha_{b,t}}{1-\alpha_{b,t}} \frac{f_{b,t}^1(\theta_b)}{f_{b,t}^0(\theta_b)} - 1}{1 - \frac{\alpha_{a,t}}{1-\alpha_{a,t}} \frac{f_{a,t}^1(\theta_a)}{f_{a,t}^0(\theta_a)}} \frac{1-\alpha_{b,t}}{1-\alpha_{a,t}}$	
`DP`	$1 - \frac{2}{\frac{\alpha_{b,t}}{1-\alpha_{b,t}} \frac{f_{b,t}^1(\theta_b)}{f_{b,t}^0(\theta_b)} + 1}$	$\left(1 - \frac{2}{\frac{\alpha_{b,t}}{1-\alpha_{b,t}} \frac{f_{b,t}^1(\theta_b)}{f_{b,t}^0(\theta_b)} + 1}\right)\left(\frac{2}{1 - \frac{\alpha_{a,t}}{1-\alpha_{a,t}} \frac{f_{a,t}^1(\theta_a)}{f_{a,t}^0(\theta_a)}} - 1\right)$	$\frac{2}{1 - \frac{\alpha_{a,t}}{1-\alpha_{a,t}} \frac{f_{a,t}^1(\theta_a)}{f_{a,t}^0(\theta_a)}} - 1$
`Simple`		$\frac{\alpha_{b,t} f_{b,t}^1(\theta_b) - (1-\alpha_{b,t}) f_{b,t}^0(\theta_b)}{(1-\alpha_{a,t}) f_{a,t}^0(\theta_a) - \alpha_{a,t} f_{a,t}^1(\theta_a)}$	

Table 4.1: The form of $\Psi_{\mathcal{C},t}(\theta_a, \theta_b)$ for $\mathcal{C} =$ `EqOpt`, `DP`, `Simple`.[5]

Note that under Assumption 7, both $\frac{\alpha_{s,t} f_{s,t}^1(\theta_s)}{(1-\alpha_{s,t}) f_{s,t}^0(\theta_s)}$ and $\alpha_{s,t} f_{s,t}^1(\theta_s) - (1-\alpha_{s,t}) f_{s,t}^0(\theta_s)$ are strictly increasing in $\theta_s \in \mathcal{T}_{s,t}$, $s \in \{a, b\}$, and $\theta_a(t) = \eta_t^{\mathcal{C}}(\theta_b(t))$ for some strictly increasing function. According to $\Psi_{\mathcal{C},t}(\theta_a, \theta_b)$ given in Table 4.1, the larger $\frac{n_a(t)}{n_b(t)}$ results in the larger $\frac{\alpha_{s,t} f_{s,t}^1(\theta_s)}{(1-\alpha_{s,t}) f_{s,t}^0(\theta_s)}$ and $\alpha_{s,t} f_{s,t}^1(\theta_s) - (1-\alpha_{s,t}) f_{s,t}^0(\theta_s)$, thus the larger $\theta_a(t)$ and $\theta_b(t)$.

The above theorem characterizes the impact of the underlying population on the one-shot decisions. Next we investigate how the one-shot decision impacts the underlying population.

4.4.2 Participation Dynamics

How a user reacts to the decision is captured by the retention dynamics (4.2) which is fully characterized by the retention rate. Below we introduce two types of (perceived) mistreatment as examples when the monotonicity condition is satisfied.

(1) User departure driven by model accuracy: Examples include discontinuing the use of products viewed as error-prone, e.g., speech recognition software, or medical diagnostic tools. In these cases, the determining factor is the classification error, i.e., users who experience low accuracy

[5]The cases represented by blank cells cannot happen. When $\mathcal{C} =$ `Simple`, the table only illustrates the result when $\delta_{a,t}, \delta_{b,t} \in \mathcal{T}_{a,t} \cap \mathcal{T}_{b,t} \neq \emptyset$.

have a higher probability of leaving the system. The retention rate at time t can be modeled as $\lambda_{s,t}(\theta_s) = \nu(L_{s,t}(\theta_s))$ for some strictly *decreasing* function $\nu(\cdot) : [0,1] \to [0,1]$.

(2) User departure driven by intra-group disparity: Participation can also be affected by intra-group disparity, that between users from the same demographic group but with different labels, i.e., $\mathcal{G}_s^y$ for $y \in \{0,1\}$. An example is in making financial assistance decisions where one expects to see more awards given to those qualified than to those unqualified. Denote by $D_{s,t}(\theta_s) = \Pr(Y_t = 1, \pi_{s,t}(X) = 1 | S = s) - \Pr(Y_t = 0, \pi_{s,t}(X_t) = 1 | S = s) = \int_{\theta_s}^{\infty} (\alpha_{s,t} f_{s,t}^1(x) - (1-\alpha_{s,t}) f_{s,t}^0(x)) dx$ as intra-group disparity of $\mathcal{G}_s$ at time t, then the retention rate can be modeled as $\lambda_{s,t}(\theta_s) = w(D_{s,t}(\theta_s))$ for some strictly *increasing* function $w(\cdot)$ mapping to $[0,1]$.

Theorem 15. *Consider the one-shot problem* (4.1) *defined in Section 4.4.1 under either* `Simple`*,* `EqOpt` *or* `DP` *criterion, and assume distributions $f_{s,t}(x) = f_s(x)$ are fixed over time. Then the one-shot problems in any two consecutive time steps, i.e., $\boldsymbol{O}_t, \boldsymbol{O}_{t+1}$, satisfy the monotonicity condition under dynamics* (4.2) *with $\lambda_s(\cdot)$ being either $\nu(L_s(\cdot))$ or $w(D_s(\cdot))$.*[6] *This implies that Theorem 12 holds and $(\theta_a(t), \theta_b(t))$ converges monotonically to a constant threshold $(\theta_a^\infty, \theta_b^\infty)$. Furthermore, $\lim_{t\to\infty} \frac{n_a(t)}{n_b(t)} = \frac{\beta_a}{\beta_b} \frac{1-\lambda_b(\theta_b^\infty)}{1-\lambda_a(\theta_a^\infty)}$.*

When distributions are fixed, the discrepancy between $\lambda_a(\theta_a(t))$ and $\lambda_b(\theta_b(t))$ increases over time as $(\theta_a(t), \theta_b(t))$ changes. The process is illustrated in Figure 4.1, where $\theta_a(t) \in [\eta^{\mathcal{C}}(\delta_b), \delta_a], \theta_b(t) \in [\delta_b, (\eta^{\mathcal{C}})^{-1}(\delta_a)]$ are constrained by the same interval $\forall t$. Left and right plots illustrate cases when $\lambda_s(\theta_s) = \nu(L_s(\theta_s))$ and $\lambda_s(\theta_s) = w(D_s(\theta_s))$ respectively.

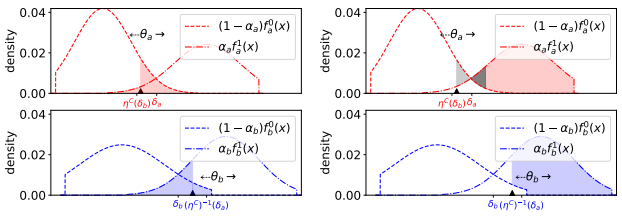


Figure 4.1: Illustration of $L_s(\theta_s)$ and $D_s(\theta_s)$ w.r.t. θ_s: Each black triangle represents the one-shot decision θ_s; size of the colored area represents the value of $L_s(\theta_s)$ (left) or $D_s(\theta_s)$ (right). Note that for the right plot, there are two gray regions and the darker one is for compensating the lighter one thus they are of the same size; the smaller gray regions result in the larger $D_a(\theta_a)$.

Note that the case considered in Theorem 15 is a special case of Theorem 13, with distributions $f_{s,t}(x) = f_s(x)$ fixed, $O_s(\theta_s) = L_s(\theta_s)$ and both dynamics $\lambda_s(\cdot) = \nu(L_s(\cdot))$ and $\lambda_s(\cdot) = w(D_s(\cdot))$ some decreasing functions of $L_s(\cdot)$.[7] In this special case we obtain the additional result of monotonic convergence of thresholds, which holds due to Assumption 7.

[6] When $f_{s,t}(x) = f_s(x)$, $\forall t$, subscript t is omitted in some notations ($\eta_t^{\mathcal{C}}, \delta_{s,t}, \lambda_{s,t}$, etc.) for simplicity.

[7] We have $D_s(\theta) = \alpha_s - L_s(\theta)$.

Once $\frac{n_a(t)}{n_b(t)}$ starts to increase, the corresponding one-shot solution $(\theta_a(t), \theta_b(t))$ also increases (Theorem 14), meaning that $\theta_a(t)$ moves closer to $\theta_a^* = \delta_a$ and $\theta_b(t)$ moves further away from $\theta_b^* = \delta_b$ (solid arrows in Figure 4.1). Consequently, $L_a(\theta_a(t))$ and $D_b(\theta_b(t))$ decrease while $L_b(\theta_b(t))$ and $D_a(\theta_a(t))$ increase. Under both dynamics, $\lambda_a(\theta_a(t))$ increases and $\lambda_b(\theta_b(t))$ decreases, resulting in the increase of $\frac{n_a(t+1)}{n_b(t+1)}$; the feedback loop becomes self-reinforcing and representation disparity worsens.

4.4.3 Impact of Decisions on Reshaping Feature Distributions

Our results so far show the potential adverse impact on group representation when imposing certain fairness criterion, while their underlying feature distributions are assumed fixed. Below we examine what happens when decisions also affect feature distributions over time, i.e., $f_{s,t}(x) = \alpha_{s,t} f^1_{s,t}(x) + (1-\alpha_{s,t}) f^0_{s,t}(x)$, which is not captured by Theorem 13. We will focus on the dynamics $\lambda_{s,t}(\theta_s) = \nu(L_{s,t}(\theta_s))$. Since $\mathcal{G}^0_s$, $\mathcal{G}^1_s$ may react differently to the same θ_s, we consider two scenarios as illustrated in Figure 4.2, which shows the change in distribution from t to $t+1$ when $\mathcal{G}^1_k$ (resp. $\mathcal{G}^0_k$) experiences the higher (resp. lower) loss at t than $t-1$: $\forall y \in \{0,1\}$,

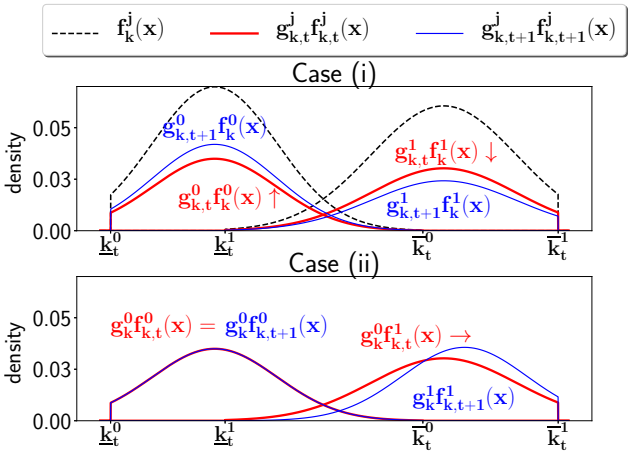


Figure 4.2: Visualization of decisions shaping feature distributions. $g^1_{k,t} = \alpha_{k,t}$, $g^0_{k,t} = 1 - \alpha_{k,t}$, where $k \in \{a, b\}$

Case (i): $f^y_{s,t}(x) = f^y_s(x)$ remain fixed but $\alpha_{s,t}$ changes over time given $\mathcal{G}^y_s$'s retention determined by its perceived loss $L^y_{s,t}$[8]. In other words, for $i \in \{0,1\}$ and $t \geq 2$ such that $L^i_{s,t}(\theta_s(t)) < L^i_{s,t-1}(\theta_s(t-1))$, we have $\alpha_{s,t+1} > \alpha_{s,t}$ if $i = 1$ and $\alpha_{s,t+1} < \alpha_{s,t}$ if $i = 0$.

Case (ii): $\alpha_{s,t} = \alpha_s$ but for subgroup $\mathcal{G}^i_s$ that is less favored by the decision over time, its members make extra effort such that $f^i_{s,t}(x)$ skews toward the direction of lowering their losses[9]. In other words, for $i \in \{0,1\}$ and $t \geq 2$ such that $L^i_{s,t}(\theta_s(t)) > L^i_{s,t-1}(\theta_s(t-1))$, we have $f^i_{s,t+1}(x) < f^i_{s,t}(x)$, $\forall x \in \mathcal{T}_s$,

[8] Here $L^1_{s,t}(\theta_s) = \int_{-\infty}^{\theta_s} f^1_{s,t}(x)dx$ and $L^0_{s,t}(\theta_s) = \int_{\theta_s}^{\infty} f^0_{s,t}(x)dx$.

[9] Suppose Assumption 7 holds for all $f^y_{s,t}(x)$ and their support does not change, then $f^1_{s,t}(x)$ and $f^0_{s,t}(x)$ overlap over $\mathcal{T}_s = [\underline{s}^1, \bar{s}^0]$, $\forall t$.

while $f^{-i}_{s,t+1}(x) = f^{-i}_{s,t}(x),\ \forall x$, where $-i := \{0,1\} \setminus \{i\}$.

In both cases, under the condition that $f_{s,t}(x)$ is relatively insensitive to the change in one-shot decisions, representation disparity can worsen and deterioration accelerates. The precise conditions are formally given in Conditions 3 and 4 in Appendix C.7, which describes the case where the change from $f_{s,t}(x)$ to $f_{s,t+1}(x)$ is sufficiently small while the change from $\frac{n_a(t)}{n_b(t)}$ to $\frac{n_a(t+1)}{n_b(t+1)}$ and the resulting decisions from $\theta_s(t)$ to $\theta_s(t+1)$ are sufficiently large. These conditions hold in scenarios when the change in feature distributions induced by the one-shot decisions is a slow process.

Theorem 16. *[Exacerbation in representation disparity can accelerate] Consider the one-shot problem defined in* (4.1) *under either* `Simple`*,* `EqOpt` *or* `DP` *fairness criterion. Let the one-shot decision, representation disparity and retention rate at time t be given by* $\theta^o_s(t)$*,* $\frac{n^o_a(t)}{n^o_b(t)}$*, and* $\lambda^o_{s,t}(\theta^o_k(t))$ *when distribution* $f_s(x)$ *is fixed* $\forall t$*. Let the same be denoted by* $\theta_s(t)$*,* $\frac{n_a(t)}{n_b(t)}$*, and* $\lambda_{s,t}(\theta_s(t))$ *when* $f_{s,t}(x)$ *changes according to either case (i) or (ii) defined above. Assume we start from the same distribution* $f_{s,1}(x) = f_s(x)$*. Under Conditions 3 and 4 in Appendix C.7, if* $\lambda^o_{a,1}(\theta^o_a(1)) = \lambda_{a,1}(\theta_a(1)) \diamond \lambda^o_{b,1}(\theta^o_b(1)) = \lambda_{b,1}(\theta_b(1))$*, then* $\frac{n_a(t+1)}{n_b(t+1)} \diamond \frac{n_a(t)}{n_b(t)}$ *(disparity worsens) and* $\frac{n_a(t+1)}{n_b(t+1)} \diamond \frac{n^o_a(t+1)}{n^o_b(t+1)}$ *(accelerates),* $\forall t$*, where* $\diamond$ *represents either* " $<$ "*or* " $>$ ".

4.4.4 Potential Mitigation & Finding the Proper Fairness Criterion From Participation Dynamics

The above results show that when the objective is to minimize the average loss over the entire population, applying commonly used and seemingly fair decisions at each time can exacerbate representation disparity over time under reasonable participation dynamics. It highlights the fact that fairness has to be defined with a good understanding of how users are affected by the algorithm, and how they may react to it. For instance, consider the dynamics with $\lambda_{s,t}(\theta_s) = \nu(L_{s,t}(\theta_s))$, then imposing `EqLos` fairness at each time step would sustain group representations, i.e., $\lim_{t\to\infty} \frac{n_a(t)}{n_b(t)} = \frac{\beta_a}{\beta_b}$, as monotonicity condition is violated under `EqLos` and we are essentially equalizing departure when equalizing loss. In contrast, under other fairness criteria the factors that are equalized do not match what drives departure, and different losses incurred to different groups cause significant change in group representation over time.

In reality the true dynamics is likely a function of a mixture of factors given the application context, and a proper fairness constraint $\mathcal{C}$ should be adopted accordingly. Below we illustrate a method for finding the proper criterion from a general dynamics model defined below when

$f_{s,t}(x) = f_s(x), \forall t$:

$$N_s(t+1) = \Lambda(N_s(t), \{\lambda_s^m(\theta_s(t))\}_{m=1}^M, \beta_s), \ \forall s \in \{a, b\}, \tag{4.4}$$

where user retention in $\mathcal{G}_s$ is driven by M different factors $\{\lambda_s^m(\theta_s(t))\}_{m=1}^M$ (e.g. accuracy, true positives, etc.) and each of them depends on decision $\theta_s(t)$. Constant β_s is the intrinsic growth rate while the actual arrivals may depend on $\lambda_s^m(\theta_s(t))$. The expected number of users at time $t+1$ depends on users at t and new users; both may be effected by $\lambda_s^m(\theta_s(t))$. This relationship is characterized by a general function Λ. Let Θ be the set of all possible decisions.

Assumption 8. *$\exists(\theta_a, \theta_b) \in \Theta \times \Theta$ such that $\forall s \in \{a, b\}$, $\hat{N}_s = \Lambda(\hat{N}_s, \{\lambda_s^m(\theta_s)\}_{m=1}^M, \beta_s)$ and $|\Lambda'(\hat{N}_s, \{\lambda_s^m(\theta_s)\}_{m=1}^M, \beta_s)| < 1$ hold for some $\hat{N}_s$, i.e., dynamics (4.4) under some decision pairs (θ_a, θ_b) have stable fixed points, where Λ' denotes the derivative of Λ with respect to N_s.*

To find the proper fairness constraint, let $\mathcal{C}$ be the set of decisions (θ_a, θ_b) that can sustain group representation. It can be found via the following optimization problem; the set of feasible solutions is guaranteed to be non-empty under Assumption 8.

$$\begin{aligned} \mathcal{C} = \arg\min_{(\theta_a, \theta_b)} \ & \left| \frac{\tilde{N}_a}{\tilde{N}_b} - \frac{\beta_a}{\beta_b} \right| \\ \text{s.t.} \ & \tilde{N}_s = \Lambda(\tilde{N}_s, \{\lambda_s^m(\theta_s)\}_{m=1}^M, \beta_s) \in \mathbb{R}_+, \theta_s \in \Theta, \forall s \in \{a, b\}. \end{aligned} \tag{4.5}$$

The idea is to first select decision pairs whose corresponding dynamics can lead to stable fixed points $(\tilde{N}_a, \tilde{N}_b)$; then among them select those that are best in sustaining group representation, which may or may not be unique. Next, we use two dynamics as examples to demonstrate how to find $\mathcal{C}$ based on dynamics.

Example 1. *[Linear first order model] is given by $N_s(t+1) = N_s(t)\lambda_s^2(\theta_s(t)) + \beta_s\lambda_s^1(\theta_s(t))$. This is a general form of dynamics (4.2) where the arrivals can also depend on the decision. When $\lambda_s^1(\theta_s(t)) = 1$, then dynamics model will be reduced to (4.2). $\tilde{N}_s = \frac{\beta_s\lambda_s^1(\theta_s)}{1-\lambda_s^2(\theta_s)}$ is the stable fixed point if $\lambda_s^2(\theta_s) < 1$ holds. Since $|\frac{\tilde{N}_a}{\tilde{N}_b} - \frac{\beta_a}{\beta_b}| = \frac{\beta_a}{\beta_b}|\frac{\lambda_a^1(\theta_a)}{\lambda_b^1(\theta_b)}\frac{1-\lambda_b^2(\theta_b)}{1-\lambda_a^2(\theta_a)} - 1|$, solution pair (θ_a, θ_b) should satisfy $\frac{\lambda_a^1(\theta_a)}{1-\lambda_a^2(\theta_a)} = \frac{\lambda_b^1(\theta_b)}{1-\lambda_b^2(\theta_b)}$. The constraint set that can sustain the group representation is given by:*

$$\mathcal{C} = \{(\theta_a, \theta_b) | (\theta_a, \theta_b) \in \Theta \times \Theta, \frac{\lambda_a^1(\theta_a)}{1-\lambda_a^2(\theta_a)} = \frac{\lambda_b^1(\theta_b)}{1-\lambda_b^2(\theta_b)}, \lambda_a^2(\theta_a) < 1, \lambda_b^2(\theta_b) < 1\}.$$

Consider the case where departure is driven by positive rate $\lambda_s^2(\theta_s) = \nu(\int_{\theta_s}^{\infty} f_s(x)dx)$ and arrival is driven by error rate $\lambda_s^1(\theta_s) = \nu((1-\alpha_s)\int_{\theta_s}^{\infty} f_s^0(x)dx + \alpha_s \int_{-\infty}^{\theta_s} f_s^1(x)dx) = \nu(L_s(\theta_s))$ where $\nu(\cdot)$ is a strictly decreasing function. This can be applied in lending scenario, where an applicant will stay as long as he/she gets the loan (positive rate) regardless of his/her qualification. Since an unqualified applicant who is issued the loan cannot repay, his/her credit score will be decreased which lowers the chance to get a loan in the future [103]. Therefore, users may decide whether to apply for a loan based on the error rate.

Example 2. *[Quadratic first order model] is given by* $N_s(t+1) = (N_s(t))^2\lambda_s^1(\theta_s(t)) + \beta_s$. $\tilde{N}_s = \frac{1}{2\lambda_s^1(\theta_s)} - \sqrt{\frac{1}{4(\lambda_s^1(\theta_s))^2} - \frac{\beta_s}{\lambda_s^1(\theta_s)}}$ *is the stable fixed point if* $\lambda_s^1(\theta_a) < \frac{1}{4\beta_s}$ *holds. Since* $\left|\frac{\tilde{N}_a}{\tilde{N}_b} - \frac{\beta_a}{\beta_b}\right| = \frac{\beta_a}{\beta_b}\left|\frac{\beta_b\lambda_b^1(\theta_b)}{\beta_a\lambda_a^1(\theta_a)}\frac{1-\sqrt{1-4\beta_a\lambda_a^1(\theta_a)}}{1-\sqrt{1-4\beta_b\lambda_b^1(\theta_b)}} - 1\right|$, *then* $\beta_a\lambda_a^1(\theta_a) = \beta_b\lambda_b^1(\theta_b)$ *should be satisfied. The constraint set that can sustain the group representation is given by*

$$\mathcal{C} = \{(\theta_a,\theta_b) | (\theta_a,\theta_b) \in \Theta\times\Theta, \beta_a\lambda_a^1(\theta_a) = \beta_b\lambda_b^1(\theta_b), \lambda_a^1(\theta_a) < \frac{1}{4\beta_a}, \lambda_b^1(\theta_b) < \frac{1}{4\beta_b}\}.$$

Sometimes guaranteeing the perfect fairness can be unrealistic and a relaxed version is preferred, in which case all pairs (θ_a,θ_b) satisfying $|\frac{\tilde{N}_a}{\tilde{N}_b} - \frac{\beta_a}{\beta_b}| \le \min\{|\frac{\tilde{N}_a}{\tilde{N}_b} - \frac{\beta_a}{\beta_b}|\} + \Delta$ constitute the Δ-fair set. An example under dynamics $N_s(t+1) = N_s(t)\lambda_s^2(\theta_s(t)) + \beta_s\lambda_s^1(\theta_s(t))$ is illustrated in Figure 4.3, where $f_s^y(x)$ is truncated normal distributed with parameters $[\sigma_a^0,\sigma_a^1,\sigma_b^0,\sigma_b^1] = [5,6,6,5]$, $[\underline{s}^0,\underline{s}^1,\overline{s}^0,\overline{s}^1] = [5,11,20,35]$, $[\mu_s^0,\mu_s^1] = [10,25]$ for $s \in \{a,b\}$. x-axis and y-axis represent θ_b and θ_a respectively. Each pair (θ_a,θ_b) has a corresponding value of $|\frac{\tilde{N}_a}{\tilde{N}_b} - \frac{\beta_a}{\beta_b}|$ measuring how well it can sustain the group representation. The colored area illustrates all the pairs such that $|\frac{\tilde{N}_a}{\tilde{N}_b} - \frac{\beta_a}{\beta_b}| \le \frac{\beta_a}{\beta_b}$. All (θ_a,θ_b) that have the same value of $|\frac{\tilde{N}_a}{\tilde{N}_b} - \frac{\beta_a}{\beta_b}| = \frac{\beta_a}{\beta_b}\epsilon$ form a curve of the same color, where the corresponding value of $\epsilon \in [0,1]$ is shown in the color bar. All curves with $\epsilon \le \Delta\frac{\beta_b}{\beta_a}$ constitute Δ-fair set (perfect fairness set is given by the deepest red curve

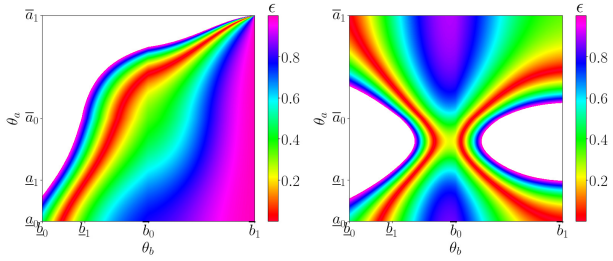
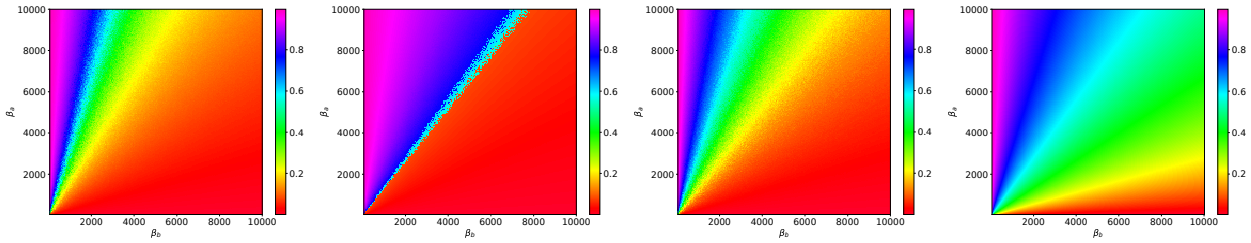


Figure 4.3: Left plot: $\lambda_s^2(\theta_s) = \nu(\int_{\theta_s}^{\infty} f_s(x)dx)$, $\lambda_s^1(\theta_s) = \nu(L_s(\theta_s))$; right plot: $\lambda_s^2(\theta_s) = \nu(L_s(\theta_s))$, $\lambda_s^1(\theta_s) = 1$, and $\nu(x) = 1-x$. Value of each pair (θ_a,θ_b) corresponds to $|\frac{\tilde{N}_a}{\tilde{N}_b} - \frac{\beta_a}{\beta_b}|$ measuring how well it can sustain the group representation. All points (θ_a,θ_b) with the same value of $|\frac{\tilde{N}_a}{\tilde{N}_b} - \frac{\beta_a}{\beta_b}| = \frac{\beta_a}{\beta_b}\epsilon$ form a curve of the same color with $\epsilon \in [0,1]$ shown in the color bar.



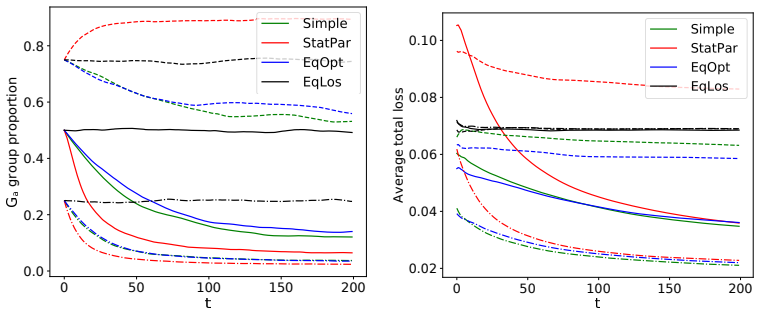
(a) `Simple` fair (b) `DP` fair (c) `EqOpt` fair (d) `EqLos` fair

Figure 4.5: Each dot in Figures 4.5a-4.5d represents the final group proportion $\lim_{t\to\infty} n_a(t)$ of one sample path under a pair of arriving rates (β_a,β_b). If the group representation is sustained, then $\lim_{t\to\infty} n_a(t) = \frac{1}{1+\beta_b/\beta_a}$ for each pair of (β_a,β_b), as shown in Figure 4.5d under `EqLos` fairness. However, under `Simple`, `DP` and `EqOpt` fairness, $\lim_{t\to\infty} n_a(t) = 1/(1+\frac{\beta_b(1-\nu(L_a(\theta_a^\infty)))}{\beta_a(1-\nu(L_b(\theta_b^\infty)))})$.

with $\epsilon = 0$).

4.5 Experiments

We first performed a set of experiments on synthetic data where every $f_s^y(x)$ follows a truncated normal distribution, the supports of $f_s^y(x), s \in \{a,b\}, y \in \{0,1\}$ are $[\underline{a}^0,\underline{a}^1,\overline{a}^0,\overline{a}^1] = [-8,5,19,35]$, $[\underline{b}^0,\underline{b}^1,\overline{b}^0,\overline{b}^1] = [-6,25,9,43]$, with the means $[\mu_a^0,\mu_a^1,\mu_b^0,\mu_b^1] = [4,20,8,27]$ and standard deviations $[\sigma_a^0,\sigma_a^1,\sigma_b^0,\sigma_b^1] = [5,6,3,6]$. The label proportions are $\alpha_a = 0.6$, $\alpha_b = 0.4$. A sequence of one-shot fair decisions are used and group representation changes over time according to dynamics (4.2) with $\lambda_s(\theta_s) = \nu(L_s(\theta_s)) = 1 - L_s(\theta_s)$.



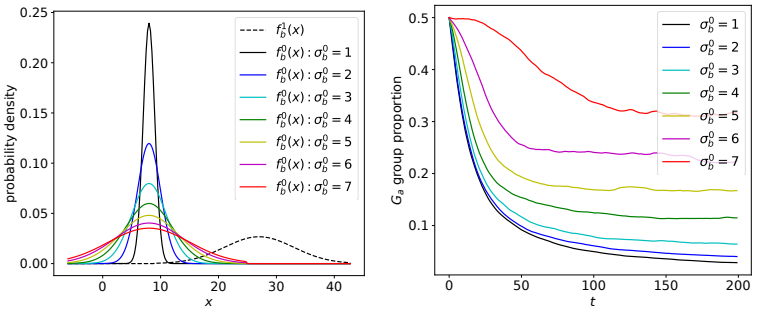
(a) Group proportion (b) Average total loss

Figure 4.4: Sample paths under different fairness criteria when $\beta_a + \beta_b = 20000$. Group proportion $n_a(t)$ and average total loss are shown in Figures 4.4a and 4.4b respectively: solid lines are for the case $\beta_a = \beta_b$, dashed lines for $\beta_a = 3\beta_b$, and dotted dashed lines for $\beta_a = \beta_b/3$.

Figure 4.4 shows sample paths of $n_a(t)$ and average total loss using one-shot fair decisions

under dynamics with $\lambda_{s,t}(\cdot) = \nu(L_{s,t}(\cdot))$. In all cases convergence is reached (we did not include the decisions $\theta_s(t)$ but convergence holds there as well). In particular, under `EqLos` fairness, the group representation is sustained throughout the horizon. By contrast, under other fairness constraints, even a "major" group (one with a larger arrival β_s) can be significantly marginalized over time (blue/green dashed line in Figure 4.4a). This occurs when the loss of the minor group happens to be smaller than that of the major group, which is determined by feature distributions of the two groups (see Figure 4.6). Whenever this is the case, the one-shot fair decision will seek to increase the minor group's proportion in order to drive down the average loss.

Figure 4.5 illustrates the final group proportion (the converged state) $\lim_{t\to\infty} n_a(t)$ as a function of the exogenous arrival sizes β_a and β_b under different fairness criteria. With the exception of `EqLos` fairness, group representation is severely skewed in the long run, with the system consisting mostly of $\mathcal{G}_b$, even for scenarios when $\mathcal{G}_a$ has larger arrival, i.e., $\beta_a > \beta_b$. Moreover, decisions under an inappropriate fairness criterion (`Simple`, `EqOpt` or `DP`) can result in poor robustness, where a minor change in β_a and β_b can result in very different representation in the long run (Figure 4.5b).



(a) Feature distributions (b) Group proportion $\beta_a = \beta_b$

Figure 4.6: Change $f_b^0(x)$ by varying $\sigma_b^0 \in \{1,2,3,4,5,6,7\}$. As σ_b^0 increases, the overlap area with $f_b^1(x)$ also increases as shown in Figure 4.6a. Figure 4.6b shows the result under `DP` fairness. Given $\theta_a(t)$, the larger σ_b^0 results in the larger $L_b(\theta_b(t))$ and thus the smaller $\mathcal{G}_b$'s retention rate.

We also consider the dynamics presented in Figure 4.3 and show the effect of $\Delta = \epsilon\frac{\beta_a}{\beta_b}$-fair decision found with method in Section 4.4.4 on $n_a(t)$. Each curve in Figure 4.7 represents a sample path under different ϵ where $(\theta_a(t), \theta_b(t))$ is from a small randomly selected subset of Δ-fair set, $\forall t$ (to model the situation where perfect fairness is not feasible) and $\beta_a = \beta_b$. We observe that fairness is always violated at the beginning in lower plot even with small ϵ. This is because the fairness set is found based on stable fixed points, which only concerns fairness in the long run.

We also trained binary classifiers over *Adult* dataset [34] by minimizing empirical loss where features are individual data points such as sex, race, and nationality, and labels are their annual income ($\geq 50k$ or $< 50k$). Since the dataset does not reflect dynamics, we employ (4.2) with $\lambda_s(\theta_s) = \nu(L_s(\theta_s))$

and $\beta_a = \beta_b$. We examine the monotonic convergence of representation disparity under `Simple`, `EqOpt` (equalized false positive/negative cost(FPC/FNC)) and `EqLos`, and consider cases where $\mathcal{G}_a$, $\mathcal{G}_b$ are distinguished by the three features mentioned above. These results are shown in Figure 4.8.

To sustain the group representation, the key point is that the fairness definition should match the factors that drive user departure and arrival. If adopt different dynamic models, different fairness criteria should be adopted. We further consider other types of dynamics and examine the performance of four fairness criteria. The results are shown in Figure 4.9. Specifically, Figure 4.9a illustrates the model where the user departure is driven by false negative rate: $N_s(t+1) = N_s(t)\nu(\mathrm{FN}_s(\theta_s(t))) + \beta_s$, with $\mathrm{FN}_s(\theta_s(t)) = \int_{\theta_s(t)}^{\infty} f_s^0(x)dx$. Under this model `EqOpt` is better at maintaining representation. Figure 4.9b illustrates the model where the users from each sub-group $\mathcal{G}_s^y$ are driven by their own perceived loss: $N_s^y(t+1) = N_s^y(t)\nu(L_s^y(\theta_s(t))) + \beta_s^y$, with $L_s^y(\theta_s)$ being false positives for $y = 0$ and false negatives for $y = 1$, arrivals $\beta_s^0 = (1-\alpha_s)\beta_s$ and $\beta_s^1 = \alpha_s\beta_s$. Under this model none of the four criteria can maintain group representation.

If $f_s^y(x)$ is unknown to the decision maker and the decision is learned from users in the system, then as users leave the system the decision can be more inaccurate and the exacerbation could potentially get more severe. In order to illustrate this, we first modify the dynamic model as $N_s(t+1) = (N_s(t) + \beta_s)\nu(L_s(\theta_s(t)))$ so that the users' arrivals are also effected by the model accuracy.[10]

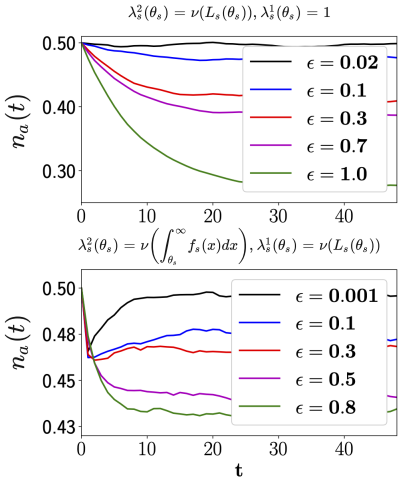


Figure 4.7: Effect of Δ-fair decisions found with proposed method.

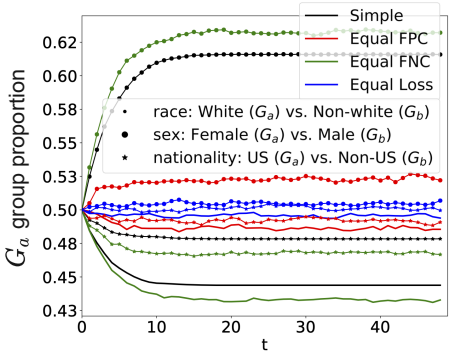
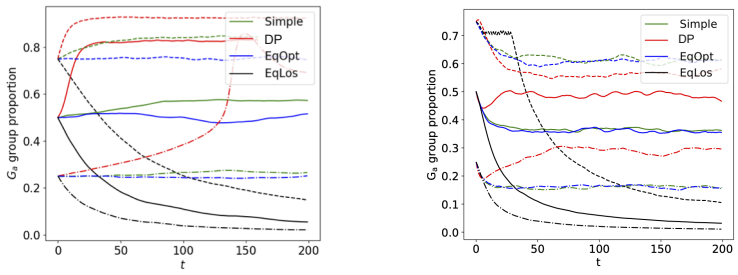


Figure 4.8: Illustration of group representation disparity using *Adult* dataset.

[10]The size of one group can decrease under this dynamics, while the size of either group is always increasing under dynamics (4.2).

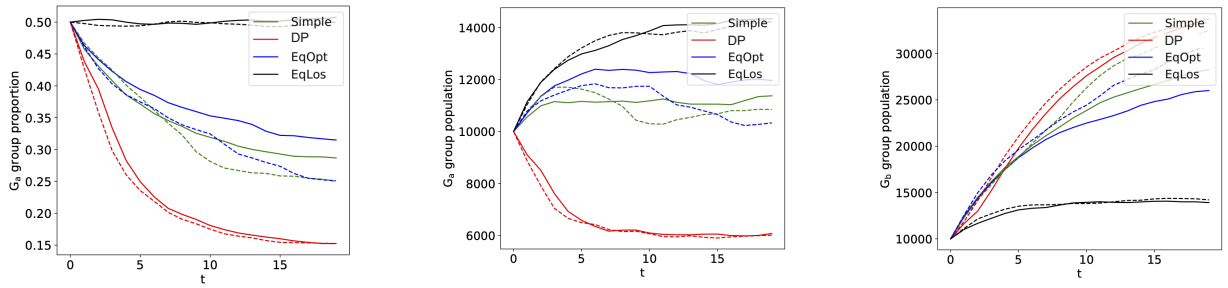
We compare the performance of two cases: *(i)* the Bayes optimal decisions are applied in every round; and *(ii)* decisions in $(t+1)$-th round are learned from the remaining users in t-th round. The empirical results are shown in Figure 4.10 where each solid curve (resp. dashed curve) is a sample path of case *(i)* (resp. case *(ii)*). Although $\beta_a = \beta_b$, $\mathcal{G}_b$ suffers a smaller loss at the beginning and starts to dominate the overall objective gradually. It results in the less and less users from $\mathcal{G}_a$ than $\mathcal{G}_b$ in the sample pool and the model trained from minority group $\mathcal{G}_a$ suffers an additional loss due to its insufficient samples. In contrast, as $\mathcal{G}_b$ becomes more dominant in the objective and its loss may be decreased compared with the case *(i)* (See Figure 4.10c). Therefore, the exacerbation in group representation disparity gets more severe (See Figure 4.10a).



(a) Users from $\mathcal{G}_s$ are driven by false negative rate

(b) Users from $\mathcal{G}_s^y$ are driven by their own perceived loss

Figure 4.9: Sample paths under different dynamic models: $\beta_a = \beta_b$ (solid curves); $\beta_a = 3\beta_b$ (dashed curves); $\beta_a = \beta_b/3$ (dotted dash curves).



(a) Group proportion

(b) $\mathcal{G}_a$'s total population

(c) $\mathcal{G}_b$'s total population

Figure 4.10: Impact of the classifier's quality: dashed curves represent the results for decisions learned from users (case *(ii)*), solid curves represent the results for Bayes optimal decisions (case *(i)*). It shows the exacerbation of group disparity get more severe under case *(ii)* for `Simple`, `EqOpt` and `DP` criteria.

CHAPTER 5

Long-Term Impact of Fairness Interventions on Group Qualification

5.1 Introduction

In Chapter 4, we focus on one type of interplay between individuals and ML system where individuals respond to ML decisions by leaving/staying ML system, and we aim to understand how group representation disparity evolves in the long run under different ML (fair) decisions. In this chapter, we study the dynamics of group qualifications [83, 104, 113, 137]. In particular, we consider scenarios where a decision maker (e.g., bank) observes individuals' features (e.g., credit scores), and makes myopic decisions (e.g., issue loans) by assessing individuals' qualifications (e.g., ability to repay) which are unknown and unobservable to the decision maker when making decisions. Individuals respond to the decisions by investing in effort to either improve or maintain their qualifications in the next time step. These actions collectively change the qualification rate of the population. Our goal is to evaluate the long-term impact of various fairness criteria and examine whether these fairness criteria mitigate or worsen the qualification disparity in the long run.

Our main contributions and findings are as follows.

1. We use a Partially Observed Markov Decision Process (POMDP) framework to model the sequential decision making and construct a qualification dynamics model to characterize the interactions between individuals and ML system (Section 5.3).
2. We analyze the equilibrium of qualification rates in different groups under a general class of fairness constraints (Section 5.4).

- We prove the existence of the equilibrium under the qualification dynamics model (Theorem 17)
- We identify the sufficient conditions under which the equilibrium of this dynamics model is unique (Theorem 18).

3. We study the long-term impact of fairness constraints on the group qualification disparity when the equilibrium is unique (Section 5.5).
   - We consider scenarios where the equality can be attained without fairness intervention (natural equality), and examine the impact of fairness constraints (Theorem 19).
   - We consider scenarios where different groups have different qualification rate at the equilibrium without fairness intervention (natural inequality), and examine the impact of fairness constraints (Theorems 20 and 21). Our findings suggest that the same fairness constraint can have opposite impacts on the equilibrium and we identify conditions under which fairness constraint can mitigate/exacerbate the inequality in the long run.
4. We explore alternative interventions that can be effective in improving qualification rates at the equilibrium and promoting equality across different groups (Section 5.6).
5. We examine our theory on synthetic Gaussian datasets and two real-world scenarios (Section 5.8). Our experiments show that our framework can help examine findings cross domains and support real-life policy making.

The remainder of this chapter is organized as follows. Section 5.2 presents related work. Section 5.3 formulates the problem. Section 5.4 conducts the equilibrium analysis. The impact of various fairness criteria on group qualification disparity is analyzed and presented in Section 5.5. The effective interventions are introduced in Section 5.6. Experiments are presented in Section 4.5. Section 5.8 concludes the chapter. All proofs can be found in Appendix D.

5.2 Related Work

Among existing works on fairness in sequential decision making problems [154], many assume that the population's feature distribution neither changes over time nor is it affected by decisions; examples include studies on handling bias in online learning [12,31,38,39,51,62,77,95] and bandits problems [8,26,73,74,97,105,117,129]. The goal of most of these work is to design algorithms that can learn near-optimal policy quickly from the sequentially arrived data and the partially observed

information, and understand the impact of imposing fairness intervention on the learned policy (e.g., total utility, learning rate, sample complexity, etc.)

However, recent studies [6, 23, 47] have shown that there exists a complex interplay between algorithmic decisions and individuals, e.g., user participation dynamics [60, 152, 153], strategic reasoning in a game [66, 83], etc., such that decision making directly leads to changes in the underlying feature distribution, which then feeds back into the decision making process. Many studies thus aim at understanding the impacts of imposing fairness constraints when decisions affect underlying feature distribution. For example, [63, 79, 83, 103] construct two-stage models where only the one-step impacts of fairness intervention on the underlying population are examined but not the long-term impacts in a sequential framework; [70, 114] focus on the fairness in reinforcement learning, of which the goal is to learn a long-run optimal policy that maximizes the cumulative rewards subject to certain fairness constraint; [60, 153] construct a user participation dynamics model where individuals respond to perceived decisions by leaving the system uniformly at random. The goal is to understand the impact of various fairness interventions on group representation.

Our work is most relevant to [66, 104, 113, 137], which study the long-term impacts of decisions on the groups' qualification states with different dynamics. In [66, 104], strategic individuals are assumed to be able to observe the current policy, based on which they can manipulate the qualification states strategically to receive better decisions. However, there is a lack of study on the influence of the sensitive attribute on dynamics and impact of fairness constraints. Besides, in many cases, the qualification states are affected by both the policy and the qualifications at the previous time step, which is considered in [113, 137]. However, they assume that the decision maker have access to qualification states and the dynamics of the qualification rates is the same in different groups, i.e.,the equally qualified people from different groups after perceiving the same decision will have the same future qualification state. In fact, the qualification states are unobservable in most cases, and the dynamics can vary across different groups. If considering such difference, the dynamics can be much more complicated such that the social equality can not be attained easily as concluded in [113, 137].

5.3 Problem Formulation

Partially Observed Markov Decision Process (POMDP). Consider two groups $\mathcal{G}_a$ and $\mathcal{G}_b$ distinguished by a sensitive *attribute* $S \in \{a, b\}$ (e.g., gender), with fractions $n_s = \Pr(S = s)$ of the popula-

tion. At time t, an *individual* with attribute $S = s$ has feature[1] $X_t = x \in \mathbb{R}$ determined by a hidden *qualification* state $Y_t = y \in \{0, 1\}$, both are time-varying. We adopt a natural assumption that an individual's attribute and current features constitute sufficient statistics, so that conditioned on these, the decision is independent of past features and decisions. This allows a decision maker to adopt a Markov policy: it makes decisions $D_t = d \in \{0, 1\}$ (reject or accept) using a policy [2] $\pi_{s,t}(x) = P_{D_t|X_t,s}(1|x, s)$ to maximize an instantaneous utility $R_t(D_t, Y_t)$, possibly subject to certain constraints. An individual is informed of the decision, and subsequently takes actions that may change the qualification Y_{t+1} and features X_{t+1}. The latter is used to drive the institute's decision at the next time step. This process is shown in Figure 5.1. Note that this model can be viewed as capturing either a randomly selected individual repeatedly going through the decision cycles, or population-wide average when all individuals are subject to the decision cycles. Thus, $\alpha_s(t) = P_{Y_t|S}(1|s)$ is the probability of an individual from $\mathcal{G}_s$ qualified at time t at the individual level, while being the *qualification rate* at the group level. One of our primary goals is to study how $\alpha_s(t)$ evolves under different (fair) policies.

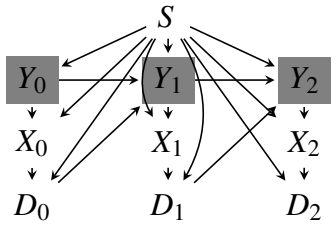


Figure 5.1: The graphical representation of our model where gray shades indicate latent variables.

Feature generation process. In many real-world scenarios, equally qualified individuals from different groups can have different features, potentially due to the different culture backgrounds and physiological differences of different demographic groups.Therefore, we consider that at time step t, given $Y_t = y$ and $S = s$, features X_t are generated by $f_s^y(x) = P_{X_t|Y_t,S}(x|y, s)$. This will be referred to as the *feature distribution* and assumed time-invariant. The convex combination $P_{X_t|S}(x|s) = \alpha_s(t) f_s^1(x) + (1 - \alpha_s(t)) f_s^0(x)$ will be referred to as the *composite feature distribution* of group $\mathcal{G}_s$ at time t.

Transition of qualification states. At time t, after receiving decision D_t, an individual takes actions such as exerting effort/investment, imitating others, etc., which results in a new qualification

[1] For simplicity of exposition, our analysis is based on one-dimensional feature space. However, the conclusions hold for high-dimensional features. This can be done by first mapping the feature space to a one-dimensional qualification profile space.

[2] We use group-dependent policies so that the optimal policies can achieve the *perfect* fairness, i.e., certain statistical measures are equalized *exactly*, which allows us to study the impact of fairness constraint *precisely*. Although using group-dependent policies might be prohibited in some scenarios (e.g., criminal justice), our qualitative conclusions are applicable to cases when two groups share the same policy, under which the *approximate* fairness is typically attained to maximize utility.

Y_{t+1}. This is modeled by a set of transitions $T_s^{yd} = P_{Y_{t+1}|Y_t,D_t,S}(1|y,d,s)$, which are time-invariant and group-dependent. These transitions characterize individuals' ability to maintain or improve its qualification. Note that we don't model individuals' strategic responses as in [66, 83], but rather use T_s^{yd} to capture the overall effect; in other words, this single quantity may encapsulate the individual's willingness to exert effort, the cost of such effort, as well as the strength of community support, etc. Specifically, T_s^{0d} (resp. T_s^{1d}) represents the probability of individuals from $\mathcal{G}_s$ who were previously unqualified (resp. qualified) became (resp. remain) qualified after receiving decision $d \in \{0,1\}$. Note that the case when feature distributions or transitions are group-independent is a special case of our formulation, i.e., by setting $f_a^y = f_b^y$ or $T_a^{yd} = T_b^{yd}$.

Fair myopic policy of an institute. A myopic policy π_t at time t aims at maximizing the instantaneous expected utility/reward $\mathcal{U}(D_t, Y_t) = \mathbb{E}[R_t(D_t, Y_t)]$, where the institute gains $u_+ > 0$ by accepting a qualified individual and incurs a cost $u_- > 0$ by accepting an unqualified individual, i.e., $R_t(D_t, Y_t) = \begin{cases} u_+, & \text{if } Y_t = 1 \text{ and } D_t = 1 \\ -u_-, & \text{if } Y_t = 0 \text{ and } D_t = 1 \\ 0, & \text{if } D_t = 0 \end{cases}$. A fair myopic policy maximizes the above utility subject to a fairness constraint $\mathcal{C}$. We focus on a set of group fairness constraints that equalize certain statistical measure between $\mathcal{G}_a$ and $\mathcal{G}_b$. A commonly studied (one-shot) fair machine learning problem is to find $(\pi_{a,t}, \pi_{b,t})$ that solves the following constrained optimization,

$$\begin{aligned} \max_{\pi_a,\pi_b} \quad & \mathcal{U}(D_t, Y_t) = n_a \mathbb{E}[R_t(D_t, Y_t)|S = a] + n_b \mathbb{E}[R_t(D_t, Y_t)|S = b] \\ \text{s.t.} \quad & \mathbb{E}_{X_t \sim \mathcal{P}_a^{\mathcal{C}}}[\pi_a(X_t)] = \mathbb{E}_{X_t \sim \mathcal{P}_b^{\mathcal{C}}}[\pi_b(X_t)] \,, \end{aligned} \tag{5.1}$$

where $\mathcal{P}_s^{\mathcal{C}}$ is some probability distribution over features X_t and specifies the fairness metric $\mathcal{C}$. Many fairness metrics including `EqOpt` and `DP` can be written in this form, i.e.,

1. Equality of Opportunity (`EqOpt`) [57]: this requires the true positive rate (TPR) to be equal, i.e., $P_{D_t|Y_t,S}(1|1,a) = P_{D_t|Y_t,S}(1|1,b)$. This is equivalent to $\mathbb{E}_{X_t|Y_t=1,S=a}[\pi_{a,t}(X_t)] = \mathbb{E}_{X_t|Y_t=1,S=b}[\pi_{b,t}(X_t)]$, i.e., $\mathcal{P}_s^{\texttt{EqOpt}}(x) = f_s^1(x)$.
2. Demographic Parity (`DP`) [11]: this requires the positive rate (PR) to be equal, i.e., $P_{D_t|S}(1|a) = P_{D_t|S}(1|b)$. This is equivalent to $\mathbb{E}_{X_t|S=a}[\pi_{a,t}(X_t)] = \mathbb{E}_{X_t|S=b}[\pi_{b,t}(X_t)]$, i.e., $\mathcal{P}_s^{\texttt{DP}}(x) = (1 - \alpha_s(t))f_s^0(x) + \alpha_s(t)f_s^1(x)$.

We focus on this class of myopic polices in this chapter, and refer to the solution to (5.1) as the

optimal policy. We further define *qualification profile*[3], $\gamma_{s,t}(x)$, the probability an individual with features x from group $\mathcal{G}_s$ is qualified at t, i.e.,

$$\gamma_{s,t}(x) = P_{Y_t|X_t,S}(1|x,s) = \frac{1}{\frac{f_s^0(x)}{f_s^1(x)}(\frac{1}{\alpha_s(t)} - 1) + 1}, \quad x \in \mathbb{R}. \tag{5.2}$$

Then the utility obtained from the group $\mathcal{G}_s$ at time step t is given by $\mathbb{E}[R_t(D_t, Y_t)|S = s] = \mathbb{E}_{X_t|S=s}[\pi_{s,t}(X_t)(\gamma_{s,t}(X_t)(u_+ + u_-) - u_-)]$. .

5.4 Evolution and Equilibrium Analysis of Qualification Rates

In this section, we first solve the one-shot optimization problem (5.1) (Section 5.4.1). We then show that under the optimal policy, there exists an equilibrium of qualification rates in the long run, and that a sufficient condition for its uniqueness is also introduced (Section 5.4.2).

5.4.1 Threshold Policies are Optimal

If an individual's qualification is observable, the optimal policy is straightforward absent of fairness constraints: accepting all qualified ones and rejecting the rest. When qualification is not observable, the institute needs to infer from observed features and accepts those most likely to be qualified. Next we show that under mild assumptions, optimal policies are in the form of threshold policies.

Assumption 9. *$f_s^y(x)$ and the CDF, $\int_{-\infty}^{x} f_s^y(z)dz$, are continuous in $x \in \mathbb{R}$, $\forall y, s$; $f_s^1(x)$ and $f_s^0(x)$ satisfy strict monotone likelihood ratio property, i.e., $\frac{f_s^1(x)}{f_s^0(x)}$ is strictly increasing in $x \in \mathbb{R}$.*

Assumption 10. *$\forall s \in \{a, b\}$, $\mathcal{P}_s^{\mathcal{C}}(x)$ is continuous in $x \in \mathbb{R}$; $\frac{P_{X|S}(x|s)}{\mathcal{P}_s^{\mathcal{C}}(x)}$ is non-decreasing in $x \in \mathbb{R}$.*

Assumption 9 says that an individual is more likely to be qualified as his/her feature value increases[4]. We show that *under Assumption 9, the optimal unconstrained policy is a threshold policy, i.e., $\forall x, t$ and $s \in \{a, b\}$, $\pi_{s,t}(x) = \mathbf{1}(x \geq \theta_s(t))$ for some $\theta_s(t) \in \mathbb{R}$.* Assumption 10 limits the types of fairness constraints, but is satisfied by many commonly used ones, including `EqOpt` and

[3] We assume the institute has perfect knowledge of $\gamma_{s,t}(x)$. In practice, this is obtained via learning/estimating $\alpha_{s,t}$ and $f_s^y(x)$ from data [71, 118].

[4] When qualification increases as the feature value x decreases, one can simply use the opposite of x.

DP. We show that *for any fairness constraint $\mathcal{C}$ satisfying Assumption 10, the optimal fair policy is a threshold policy*, which is consistent with Theorem 3.2 in [29]. Moreover, under Assumptions 9 and 10, a threshold as a function of qualification rates, $\theta_s(t) = \theta_s(\alpha_a(t), \alpha_b(t))$, is continuous and non-increasing in $\alpha_a(t)$ and $\alpha_b(t)$. In the next Lemma 7, we further characterize these optimal (fair) thresholds in the optimal (fair) policies.

Lemma 7 (Optimal (fair) threshold)**.** *Let $(\gamma_a(x), \gamma_b(x))$ be a pair of qualification profiles for groups $\mathcal{G}_a$ and $\mathcal{G}_b$ at t. Let threshold pairs $(\theta_a^{UN}, \theta_b^{UN})$ and $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$ be the unconstrained and fair optimal thresholds under constraint $\mathcal{C}$, respectively. Then we have $\gamma_a(\theta_a^{UN}) = \gamma_b(\theta_b^{UN}) = \frac{u_-}{u_+ + u_-}$ and*

$$n_a\Big(\gamma_a(\theta_a^{\mathcal{C}}) - \frac{u_-}{u_+ + u_-}\Big)\frac{P_{X|S}(\theta_a^{\mathcal{C}}|a)}{\mathcal{P}_a^{\mathcal{C}}(\theta_a^{\mathcal{C}})} + n_b\Big(\gamma_b(\theta_b^{\mathcal{C}}) - \frac{u_-}{u_+ + u_-}\Big)\frac{P_{X|S}(\theta_b^{\mathcal{C}}|b)}{\mathcal{P}_b^{\mathcal{C}}(\theta_b^{\mathcal{C}})} = 0. \tag{5.3}$$

Here we have removed the subscript t since the thresholds are not t-dependent; they only depend on current qualification rates. The solution to Eqn. (5.3) is the threshold pair $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$ that satisfies the fairness constraint $\int_{\theta_a^{\mathcal{C}}}^{\infty} \mathcal{P}_a^{\mathcal{C}}(x)dx = \int_{\theta_b^{\mathcal{C}}}^{\infty} \mathcal{P}_b^{\mathcal{C}}(x)dx$ in Eqn. (5.1) while maximizing the expected utility $\mathcal{U}(D, Y)$ at time t. Under DP and EqOpt fairness, Eqn. (5.3) can be reduced to

$$n_a\gamma_a(\theta_a^{\texttt{DP}}) + n_b\gamma_b(\theta_b^{\texttt{DP}}) = \frac{u_-}{u_+ + u_-}; \qquad \frac{n_a\alpha_a}{\gamma_a(\theta_a^{\texttt{EqOpt}})} + \frac{n_b\alpha_b}{\gamma_b(\theta_b^{\texttt{EqOpt}})} = \frac{n_a\alpha_a}{\frac{u_-}{u_+ + u_-}} + \frac{n_b\alpha_b}{\frac{u_-}{u_+ + u_-}}.$$

Lemma 7 also indicates the relation between the unconstrained and fair optimal polices, e.g., a group's qualification profile evaluated at the unconstrained threshold is the same as the weighted combination of two groups' qualification profiles evaluated at their corresponding fair thresholds under DP.

5.4.2 Evolution and Equilibrium Analysis

We next examine what happens as the institute repeatedly makes decisions based on the optimal (fair) policies derived in Section 5.4.1, while individuals react by taking actions to affect their future qualifications. We say the qualification rate of $\mathcal{G}_s$ is at an *equilibrium* if $\alpha_s(t+1) = \alpha_s(t), \forall t \geq t_o$ for some t_o, or equivalently, if $\lim_{t\to\infty} \alpha_s(t) = \alpha_s$ is well-defined for some $\alpha_s \in [0, 1]$. Analyzing equilibrium helps us understand the property of the population in the long-run. We begin by

characterizing the dynamics of qualification rates $\alpha_s(t)$ under policy $\pi_{s,t}$ as follows:

$$\alpha_s(t+1) = g_s^0(\alpha_a(t), \alpha_b(t)) \cdot (1 - \alpha_s(t)) + g_s^1(\alpha_a(t), \alpha_b(t)) \cdot \alpha_s(t), \quad s \in \{a, b\}\,, \tag{5.4}$$

where $g_s^y(\alpha_a(t), \alpha_b(t)) = \mathbb{E}_{X_t|Y_t=y,S=s}\left[(1-\pi_{s,t}(X_t))T_s^{y0} + \pi_{s,t}(X_t)T_s^{y1}\right]$ depends on qualification rates $\alpha_a(t), \alpha_b(t)$ through policy $\pi_{s,t}$. When $\pi_{s,t}(x) = \mathbf{1}(x \geq \theta_s(t))$, this reduces to $g_s^y(\alpha_a(t), \alpha_b(t)) = T_s^{y0}\int_{-\infty}^{\theta_s(t)} f_s^y(x)dx + T_s^{y1}\int_{\theta_s(t)}^{\infty} f_s^y(x)dx$. Denote $g_s^y(\alpha_a(t), \alpha_b(t)) := g_s^y(\theta_s(\alpha_a(t), \alpha_b(t)))$, $y \in \{0, 1\}$.

Dynamics (5.4) says that the qualified people at each time consists of two parts: the qualified people in the previous time step remain being qualified, and those who were unqualified in the previous time step change to become qualified.

Theorem 17 below shows that for any transition and any threshold policy that are continuous in qualification rates, the above dynamical system always has at least one equilibrium.

Theorem 17 (Existence of equilibrium)**.** *Consider a dynamics* (5.4) *with a threshold policy* $\theta_s(\alpha_a, \alpha_b)$ *that is continuous in* α_a *and* α_b*.* $\forall\, T_s^{yd} \in (0, 1)$*, there exists at least one equilibrium* $(\widehat{\alpha}_a, \widehat{\alpha}_b)$*.*

While an equilibrium exists under any set of transitions, its specific property (e.g., quantity, value, etc.) highly depends on transition probabilities which specify different user dynamics.

We focus on two scenarios given in the condition below.

Condition 1. $\forall s \in \{a, b\}$*,*

a) $T_s^{01} \leq T_s^{00}$ *and* $T_s^{11} \leq T_s^{10}$*;* *b)* $T_s^{01} \geq T_s^{00}$ *and* $T_s^{11} \geq T_s^{10}$*.*

As mentioned, transitions T_s^{yd} characterize the ability of individuals from $\mathcal{G}_s$ to maintain/improve their future qualifications, this value summarizes individual's behaviors. On one hand, an accepted individual may feel less motivated to remain qualified (if it was) or become qualified (if it was not). On the other hand, the accepted individual may have access to better resources or feel more inspired to remain or become qualified. These competing factors (referred to later as the "lack of motivation" effect and the "leg-up" effect, respectively) may work simultaneously, and the net effect can be context dependent. Condition 1a) (resp. Condition 1b)) suggests that the first (resp. second) effect is dominant for both qualified and unqualified individuals. There are two other combinations: c) $T_s^{01} \geq T_s^{00}$ and $T_s^{11} \leq T_s^{10}$; d) $T_s^{01} \leq T_s^{00}$ and $T_s^{11} \geq T_s^{10}$, under which the qualified and unqualified are dominant by different effects. These cases incur more uncertainty; slight changes in feature distributions or transitions may result in opposite conclusions. More discussions are in Section 5.7.

Given the existence of an equilibrium, Theorem 18 further introduces sufficient conditions for it to be unique. Based on the unique equilibrium, we can evaluate and compare policies (Section 5.5), and design effective interventions to promote long-term equality and/or the overall qualifications (Section 5.6).

Theorem 18 (Uniqueness of equilibrium)**.** *Consider a decision-making system with dynamics* (5.4) *and either unconstrained or fair optimal threshold policy. Let* $h_s(\theta_s(\alpha_a,\alpha_b)) = \frac{1-g_s^1(\theta_s(\alpha_a,\alpha_b))}{g_s^0(\theta_s(\alpha_a,\alpha_b))}, s \in \{a,b\}$*. Under Assumptions 9 and 10, a sufficient condition for* (5.4) *to have a unique equilibrium is as follows,* $\forall s \in \{a,b\}$*:*

1. *Under Condition 1a),* $\left|\frac{\partial h_s(\theta_s(\alpha_a,\alpha_b))}{\partial \alpha_{-s}}\right| < 1$*,* $\forall \alpha_s \in [0,1]$*, where* $-s := \{a,b\} \setminus s$*;*
2. *Under Condition 1b),* $\left|\frac{\partial h_s(\theta_s(\alpha_a,\alpha_b))}{\partial \alpha_{-s}}\right| < 1$ *and* $\left|\frac{\partial h_s(\theta_s(\alpha_a,\alpha_b))}{\partial \alpha_s}\right| < 1, \forall \alpha_a, \alpha_b \in [0,1]$*.*

These sufficient conditions can further be satisfied if for the qualified ($y = 1$) and the unqualified ($y = 0$), transitions T_s^{y1} and T_s^{y0} are sufficiently close, i.e., policies have limited influence on the qualification dynamics. This is stated formally as follows.

Corollary 2. *For any feature distributions* $\{f_s^y(x)\}_{s,y}$*, suppose* $\left|\frac{\partial \mathbb{F}_s^y(\theta_s(\alpha_a,\alpha_b))}{\partial \alpha_u}\right| \leq M^y$ *holds for some constant* $M^y \in [0,\infty)$*,* $\forall y \in \{0,1\}, \forall u \in \{a,b\}$*. Under either Condition 1a) or 1b),* $\exists \epsilon_s^y > 0$ *such that for any transitions that satisfy* $|T_s^{y1} - T_s^{y0}| < \epsilon_s^y$*,* $s \in \{a,b\}, y \in \{0,1\}$*, the corresponding dynamics system has a unique equilibrium.*

It is worth noting that the conditions of Theorem 18 only guarantee uniqueness of equilibrium but not stability, i.e., it is possible that the qualification rates oscillate and don't converge under this discrete-time dynamics (see examples on COMPAS data in Section 5.8). The uniqueness can be guaranteed and further attained if the dynamics (5.4) satisfies L-Lipschitz condition with $L < 1$. However, Lipschitz condition is relatively stronger than the condition in Theorem 18 (see the comparison in Section 5.7).

Figure 5.2 illustrates trajectories of qualification rates $(\alpha_a(t), \alpha_b(t))$ and the equilibrium for a Gaussian case under Condition 1b). Let $g_s^y := g_s^y(\theta_s(\alpha_a,\alpha_b))$, the points $(\alpha_a,\ \alpha_b)$ on the red, and blue dashed curves satisfy $\alpha_b = g_b^0(1-\alpha_b) + g_b^1\alpha_b$ and $\alpha_a = g_a^0(1-\alpha_a) + g_a^1\alpha_a$, respectively. Their intersection (black star) is the equilibrium $(\widehat{\alpha}_a, \widehat{\alpha}_b)$. The sufficient conditions in Theorem 18 guarantee these two curves have only one intersection. Moreover, observe that these two curves split the plane $\{(\alpha_a,\alpha_b) : \alpha_a \in [0,1], \alpha_b \in [0,1]\}$ into four parts, which can be used for determining how $(\alpha_a(t), \alpha_b(t))$ will change at t. For example, if $(\alpha_a(t), \alpha_b(t))$ falls into the left side of the blue

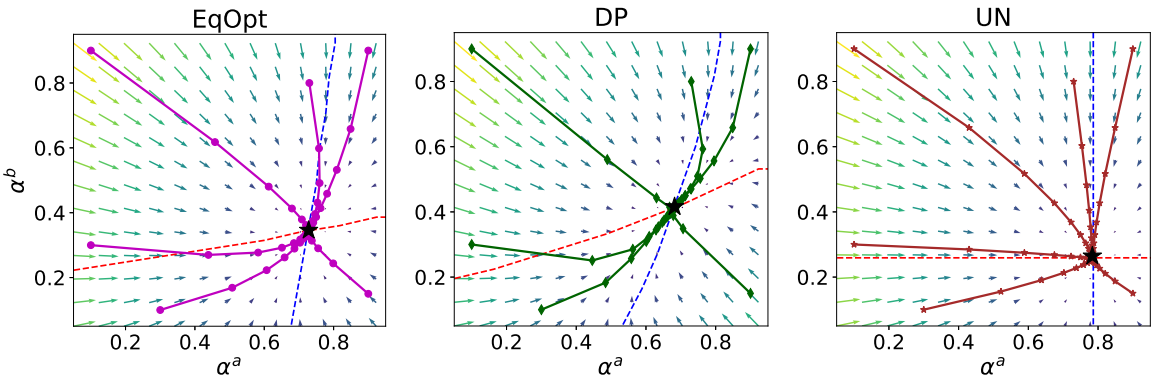


Figure 5.2: Illustration of $\{(\alpha_a(t), \alpha_b(t))\}_t$ for a Gaussian case under `EqOpt`, `DP`, `UN` optimal policies: $u_+ = u_-$, $n_a = n_b$, $f_s^y(x)$ is Gaussian distributed with mean μ_s^y and variance σ_s^2, where $[\mu_a^0, \mu_a^1, \mu_b^0, \mu_b^1] = [-5, 5, -5, 5]$, $[\sigma_a, \sigma_b] = [5, 5]$, $[T_a^{00}, T_a^{01}, T_a^{10}, T_a^{11}] = [0.4, 0.5, 0.5, 0.9]$, $[T_b^{00}, T_b^{01}, T_b^{10}, T_b^{11}] = [0.1, 0.5, 0.5, 0.7]$. Each plot shows 6 sample paths with each circle/diamond/star representing one pair of $(\alpha_a(t), \alpha_b(t))$.

dashed curve, then $\alpha_a(t+1) > \alpha_a(t)$; if $(\alpha_a(t), \alpha_b(t))$ falls into the lower side of the red dashed curve, then $\alpha_b(t+1) > \alpha_b(t)$.

5.5 The Long-Term Impact of Fairness Constraints

In this section, we analyze the long-term impact of imposing fairness constraints on the equality of group qualification. We will do so in the presence of *natural equality (and inequality)* [113] where equitable equilibria are attained naturally without imposing additional constraints (in our context, this means attaining $\widehat{\alpha}_a^{\text{UN}} = \widehat{\alpha}_b^{\text{UN}}$ using unconstrained polices).

Although there may exist multiple equilibria, in this section we will assume the conditions in Theorem 18 hold under Assumption 9 and 10 and limit ourselves to the unique equilibrium cases under `DP` and `EqOpt`, thereby providing a theoretical foundation and an illustration of how their long-term impact can be compared. As shown below, these short-term fairness interventions may not necessarily promote long-term equity, and their impact can be sensitive to feature distributions and transitions. A small change in either can lead to contrarian results, suggesting the importance of understanding the underlying population.

Long-term impact on natural equality. When there is natural equality, an unconstrained optimal policy will result in two groups converging to the same qualification rate, thus rendering

fairness constraints is unnecessary. The interesting question here is whether applying a fairness constraint can disrupt the equality. The theorem below shows that the `DP` and `EqOpt` fairness will do harm if the feature distributions are different.

Theorem 19. *For any feature distribution $f_s^y(x)$ and $\forall \alpha^{UN} \in (0,1)$, there exist transitions $\{T_s^{yd}\}_{y,d,s}$ satisfying either Condition 1a) or Condition 1b) such that $\widehat{\alpha}_a^{UN} = \widehat{\alpha}_b^{UN} = \alpha^{UN}$. In this case, if $f_a^y(x) \neq f_b^y(x)$ (resp. $f_a^y(x) = f_b^y(x)$), then imposing either $\mathcal{C}$ = `DP` or `EqOpt` fair optimal policies will violate (resp. maintain) equality, i.e., $\widehat{\alpha}_a^{\mathcal{C}} \neq \widehat{\alpha}_b^{\mathcal{C}}$ (resp. $\widehat{\alpha}_a^{\mathcal{C}} = \widehat{\alpha}_b^{\mathcal{C}}$).*

Theorem 19 shows that $\forall \alpha^{\text{UN}} \in (0,1)$, there exists model parameters under which α^{UN} is the equilibrium and natural equality is attained. Also, natural equality is not disrupted by imposing either fairness constraint when feature distributions are the same across different groups (referred to as *demographic-invariant* below). However, imposing either constraint will lead to unequal outcomes if feature distributions are diverse across groups (referred to as *demographic-variant* below), which is more likely to happen in reality. Thus, in these natural equality cases, imposing fairness will often do harm.

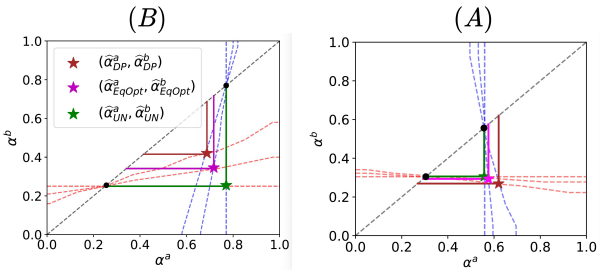
Long-term impact on natural inequality. Natural inequality, i.e., $\widehat{\alpha}_a^{\text{UN}} \neq \widehat{\alpha}_b^{\text{UN}}$, is more common than natural equality which only occurs under specific model parameters. This difference in qualification rates at equilibrium typically stems from the fact that either feature distributions or transitions or both are different across different groups. Thus, below we study the impact of imposing fairness by considering these two sources of inequality separately, and we aim to examine whether fairness constraints can address the inequality caused by each. Let *disadvantaged group* be the group with a lower qualification rate at equilibrium.

Demographic-invariant feature distribution with demographic-variant transition. In this case, we have the same feature distributions but different transitions in different groups, i.e., $f_s^y = f_b^y$, $T_a^{yd} \neq T_b^{yd}$. A real-world example is college admission based on ACT/SAT scores: given the same qualification state, score distributions may be the same regardless of the applicant's socio-economic status, but the economically advantaged may be able to afford more investments and effort to improve their score after a rejection.

Theorem 20. *Under Condition 1a), `DP` and `EqOpt` fairness exacerbate inequality, i.e., $|\widehat{\alpha}_a^{\mathcal{C}} - \widehat{\alpha}_b^{\mathcal{C}}| \geq |\widehat{\alpha}_a^{UN} - \widehat{\alpha}_b^{UN}|$; under Condition 1b), `DP` and `EqOpt` fairness mitigate inequality, i.e., $|\widehat{\alpha}_a^{\mathcal{C}} - \widehat{\alpha}_b^{\mathcal{C}}| \leq |\widehat{\alpha}_a^{UN} - \widehat{\alpha}_b^{UN}|$. Moreover, the disadvantaged group remains disadvantaged in both cases, i.e., $(\widehat{\alpha}_a^{UN} - \widehat{\alpha}_b^{UN})(\widehat{\alpha}_a^{\mathcal{C}} - \widehat{\alpha}_b^{\mathcal{C}}) \geq 0$.*

This theorem shows that imposing fairness only helps when the "leg-up" effect is more prominent than the "lack of motivation" effect; alternatively, this suggests that when the "lack of motivation" effect is dominant, imposing fairness should be accompanied by other support structure to dampen this effect (e.g., by helping those accepted to become or remain qualified).

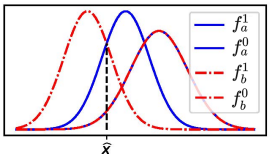
Theorem 20 is illustrated in the plot to the right, where transitions satisfy Condition 1a)-b) and $f_a^y(x) = f_b^y(x)$ is Gaussian distributed. Each plot includes 3 pairs of red/blue dashed curves corresponding to 3 policies (`EqOpt`, `DP`, `UN`). Points (α_a, α_b) on these curves satisfy $\alpha_b = g_b^0(\alpha_a, \alpha_b) \cdot (1-\alpha_b) + g_b^1(\alpha_a, \alpha_b) \cdot \alpha_b$ and $\alpha_a = g_a^0(\alpha_a, \alpha_b) \cdot (1-\alpha_a) + g_a^1(\alpha_a, \alpha_b) \cdot \alpha_a$, respectively. Each intersection (colored star) is an equilibrium $(\widehat{\alpha}_a^{\mathcal{C}}, \widehat{\alpha}_b^{\mathcal{C}})$; the length of colored segments represents $|\widehat{\alpha}_a^{\mathcal{C}} - \widehat{\alpha}_b^{\mathcal{C}}|$. The black circle is the intersection of all three blue/red curves.



Demographic-variant feature distribution with demographic-invariant transition. In this case, we have the same transitions and different feature distributions in different groups, i.e., $f_a^y \neq f_b^y$, $T_a^{yd} = T_b^{yd}$. In the same example of college admission this is a case where the ACT/SAT scores are biased against a certain group but there is no difference in how different groups react to the decision. Here, we will focus on a class of feature distributions where those qualified have the same feature distribution regardless of group membership, while those unqualified from $\mathcal{G}_b$ are more likely to have lower features than those unqualified from $\mathcal{G}_a$. This is given in the condition below.

Condition 2. *$f_s^y(x)$ is continuous in $x \in \mathbb{R}$; $f_a^1(x) = f_b^1(x), \forall x \in \mathbb{R}$; $f_a^0(x)$ and $f_b^0(x)$ satisfy strict monotone likelihood ratio property, i.e., $\frac{f_a^0(x)}{f_b^0(x)}$ is strict increasing in $x \in \mathbb{R}$.*

Condition 2 also implies that $\int_{-\infty}^{x} f_b^0(z)dz \geq \int_{-\infty}^{x} f_a^0(z)dz, \forall x \in \mathbb{R}$. Let $\widehat{x}$ be defined such that $f_b^0(\widehat{x}) = f_a^0(\widehat{x})$ holds, which is unique. An example satisfying Condition 2 is shown on the right.



Theorem 21. *Under Condition 1b) and Condition 2, if $\frac{u_+}{u_-} \geq \frac{f_s^0(\widehat{x})}{f_s^1(\widehat{x})} \frac{1-T^{10}}{T^{00}}$, we have*

- *$\widehat{\alpha}_a^{UN} > \widehat{\alpha}_b^{UN}$ and $\widehat{\alpha}_a^{UN} - \widehat{\alpha}_b^{UN} > \widehat{\alpha}_a^{EqOpt} - \widehat{\alpha}_b^{EqOpt} \geq 0$ hold, i.e., `EqOpt` fairness always mitigates inequality and the disadvantaged group $\mathcal{G}_b$ remains disadvantaged.*
- *`DP` fairness may either (1) mitigate inequality, i.e., $\widehat{\alpha}_a^{UN} - \widehat{\alpha}_b^{UN} > \widehat{\alpha}_a^{DP} - \widehat{\alpha}_b^{DP} \geq 0$; or (2) flip the disadvantaged group from $\mathcal{G}_b$ to $\mathcal{G}_a$, i.e., $\widehat{\alpha}_b^{DP} \geq \widehat{\alpha}_a^{DP}$.*

Because $\mathcal{G}_a$ and $\mathcal{G}_b$ only differ in $f_s^0(x)$, the condition in Theorem 21 ensures at least one group has enough unqualified people to be accepted and can be satisfied if benefit u_+ is sufficiently larger than cost u_-. We see that in this case the comparison is much more complex depending on the model parameters.

5.6 Effective Interventions

As discussed, imposing static fairness constraints is not always a valid intervention in terms of its long-term impact. In some cases it reinforces existing disparity; even when it could work, doing it right can be very hard due to its sensitivity to problem parameters. In this section, we present several alternative interventions that can be more effective in inducing more equitable outcomes or improving overall qualification rates in the long run. We shall assume that the sufficient conditions of Theorem 18 hold under Assumptions 9 and 10 so that the equilibrium is unique.

Policy intervention. In many instances, preserving static fairness at each time t is important, for short-term violations may result in costly lawsuits [1]. Proposition 3 below shows that using *sub-optimal* fair policies instead of the optimal ones can improve overall qualification in the long run.

Proposition 3. *Let $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$, $(\theta_a^{\mathcal{C}'}, \theta_b^{\mathcal{C}'})$ be thresholds satisfying fairness constraint $\mathcal{C}$ under the optimal and an alternative policy, respectively. Let $(\widehat{\alpha}_a^{\mathcal{C}}, \widehat{\alpha}_b^{\mathcal{C}})$, $(\widehat{\alpha}_a^{\mathcal{C}'}, \widehat{\alpha}_b^{\mathcal{C}'})$ be the corresponding equilibrium.*

- *If $\theta_s^{\mathcal{C}'}(\alpha_a, \alpha_b) > \theta_s^{\mathcal{C}}(\alpha_a, \alpha_b)$, $\forall \alpha_s \in [0,1]$ under Condition 1a), then $\widehat{\alpha}_s^{\mathcal{C}'} > \widehat{\alpha}_s^{\mathcal{C}}$, $\forall s \in \{a,b\}$;*
- *If $\theta_s^{\mathcal{C}'}(\alpha_a, \alpha_b) < \theta_s^{\mathcal{C}}(\alpha_a, \alpha_b)$, $\forall \alpha_s \in [0,1]$ under Condition 1b), then $\widehat{\alpha}_s^{\mathcal{C}'} > \widehat{\alpha}_s^{\mathcal{C}}$, $\forall s \in \{a,b\}$.*

Note that the sacrifice is in instantaneous utility, not necessarily in total utility in the long run (see an example in proof of Proposition 3, Appendix D). If static fairness need not be maintained at all times, then we can employ separate policies for each group, and Proposition 4 below shows that under certain conditions on transitions, threshold policies leading to equitable equilibrium always exist.

Proposition 4. *Let $\mathcal{I}_s = [\frac{1-\max\{T_s^{11}, T_s^{10}\}}{\max\{T_s^{01}, T_s^{00}\}}, \frac{1-\min\{T_s^{11}, T_s^{10}\}}{\min\{T_s^{01}, T_s^{00}\}}]$, $s \in \{a,b\}$. Under Condition 1a) or 1b), if $\mathcal{I}_a \cap \mathcal{I}_b \neq \emptyset$, then $\forall \widehat{\alpha} \in \mathcal{I}_a \cap \mathcal{I}_b$, there exist threshold policies $\theta_s(\alpha_s)$, $\forall \alpha_s \in [0,1]$, under which $\alpha_s(t) \to \widehat{\alpha}, \forall s \in \{a,b\}$, i.e., equitable equilibrium is attained; if $\mathcal{I}_a \cap \mathcal{I}_b = \emptyset$, then there is no threshold policy that can result in equitable equilibrium.*

Proposition 4 also indicates that when two groups' transitions are significantly different, manipulating policies cannot achieve equality. In this case, the following intervention can be considered.

Transition Intervention. Another intervention is to alter the value of transitions, e.g., by establishing support for both the accepted and rejected. Proposition 5 shows that the qualification rate $\widehat{\alpha}_s$ at equilibrium can be improved by enhancing individuals' ability to maintain/improve qualification, which is consistent with the empirical findings in loan repayment [52, 64, 119] and labor markets [43].

Proposition 5. *$\forall s \in \{a,b\}$, increasing any transition probability T_s^{yd}, $d \in \{0,1\}$, $y \in \{0,1\}$ always increases the value of equilibrium qualification rates $\widehat{\alpha}_s$.*

5.7 Discussion

Transitions under Condition 1c) or 1d). This chapter mainly focus on transitions satisfying Condition 1a) and 1b). As mentioned in Section 5.4.2, there are the other two combinations: c) $T_s^{01} \geq T_s^{00}$ and $T_s^{11} \leq T_s^{10}$; d) $T_s^{01} \leq T_s^{00}$ and $T_s^{11} \geq T_s^{10}$, in which there is more uncertainty when conducting equilibrium analysis. The slight changes in the feature distributions or the values of transitions may change conclusions significantly.

Because the system has equilibrium if $\alpha_s(t) = \alpha_s(t+1)$ holds, i.e., there is solution to $\alpha_s = g_s^0(\alpha_a,\alpha_b)\cdot(1-\alpha_s) + g_s^1(\alpha_a,\alpha_b)\cdot\alpha_s, \forall s \in \{a,b\}$. Re-organize, it requires $\frac{1}{\alpha_s} - 1 = \frac{1-g_s^1(\theta_s(\alpha_a,\alpha_b))}{g_s^0(\theta_s(\alpha_a,\alpha_b))}$, $\forall s \in \{a,b\}$. Let cumulative density function of $f_s^y(x)$ be denoted as $\mathbb{F}_s^y(x) = \int_{-\infty}^{x} f_s^y(z)dz$. Since

$$\frac{1-g_s^1(\theta_s(\alpha_a,\alpha_b))}{g_s^0(\theta_s(\alpha_a,\alpha_b))} = \frac{1-(T_s^{10}\mathbb{F}_s^1(\theta_s(\alpha_a,\alpha_b)) + T_s^{11}(1-\mathbb{F}_s^1(\theta_s(\alpha_a,\alpha_b))))}{T_s^{00}\mathbb{F}_s^0(\theta_s(\alpha_a,\alpha_b)) + T_s^{01}(1-\mathbb{F}_s^0(\theta_s(\alpha_a,\alpha_b)))}.$$

Under optimal (fair) policies and Condition 1a) or 1b), $\frac{1-g_s^1(\theta_s(\alpha_a,\alpha_b))}{g_s^0(\theta_s(\alpha_a,\alpha_b))}$ is guaranteed to be either decreasing or increasing in α_s. This monotonicity is critical to determine the properties (e.g., uniqueness, quantity, value, etc.) of the consequent equilibrium $(\widehat{\alpha}_a^{\mathcal{C}}, \widehat{\alpha}_b^{\mathcal{C}})$ so that impacts of different fairness can be compared. In contrast, under Condition 1c) or 1d), $\frac{1-g_s^1(\theta_s(\alpha_a,\alpha_b))}{g_s^0(\theta_s(\alpha_a,\alpha_b))}$ is no longer monotonic, and its intersection with function $\frac{1}{\alpha_s} - 1$, i.e., equilibrium, is thus hard to characterize. As a consequence, the impacts of different fairness constraints cannot be compared in general.

Comparison between sufficient conditions in Theorem 18 and Lipschitz condition. Let a pair of qualification rats of $\mathcal{G}_a, \mathcal{G}_b$ be noted as $\boldsymbol{\alpha}_t = (\alpha_a(t), \alpha_b(t)) \in [0,1] \times [0,1]$, and let mapping $\Phi : [0,1] \times [0,1] \to [0,1] \times [0,1]$ be defined such that dynamical system (5.4) can be written as $\boldsymbol{\alpha}_{t+1} = \Phi(\boldsymbol{\alpha}_t)$. Then this dynamical system has an equilibrium $\widehat{\alpha}$ if $\Phi(\widehat{\alpha}) = \widehat{\alpha}$. According to Banach Fixed Point Theorem, such equilibrium exists and is unique if the mapping Φ satisfies L-Lipschitz condition with $L < 1$, i.e., Φ is a contraction mapping. Specifically, $d(\Phi(\boldsymbol{\alpha}_0), \Phi(\boldsymbol{\alpha}_1)) \leq L d(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)$ for some distance function d and Lipschitz constant $L < 1$.

While Lipschitz condition also ensures the uniqueness of equilibrium, the sufficient conditions given in Theorem 18 are weaker. Use unconstrained optimal policies as an example, in this case dynamics of two groups can be decoupled because threshold $\theta_s(\alpha_a, \alpha_b)$ used in $\mathcal{G}_s$ is independent of qualification of the other group α_{-s}. Therefore, sufficient condition $|\frac{\partial h_s(\theta_s(\alpha_a, \alpha_b))}{\partial \alpha_{-s}}| = 0 < 1$ under Condition 1a) always holds. In contrast, for dynamics of $\mathcal{G}_s$ after decoupling $\alpha_s(t+1) = \Phi_s(\alpha_s(t)) = g_s^0(\theta_s(\alpha_s(t)))(1 - \alpha_s(t)) + g_s^1(\theta_s(\alpha_s(t)))\alpha_s(t)$, Φ_s is not necessarily a contraction mapping.

Although sufficient conditions in Theorem 18 are weaker, they do not guarantee the stability of the equilibrium. In contrast, Lipschitz condition with $L < 1$ ensures the unique equilibrium is also stable, i.e., we have $(\alpha_a(t), \alpha_b(t)) \to (\widehat{\alpha}_a, \widehat{\alpha}_b)$ given an arbitrary initial state $(\alpha_a(0), \alpha_b(0))$.

5.8 Experiments

We conducted experiments on both Gaussian synthetic datasets and real-world datasets (FICO credit scores and COMPAS data). These are static, one-shot datasets, which we use to create a simulated dynamic setting as detailed below.

Gaussian synthetic data. We first verify the conclusions in Sections 5.4 and 5.5 using the synthetic data, where $f_s^y(x)$ is Gaussian distributed with mean μ_s^y and variance σ_s^2.

Table 5.1 and 5.2 illustrate the impacts of `EqOpt` and `DP` fairness on the equilibrium, where each column shows the value of $\widehat{\alpha}_a^{\mathcal{C}} - \widehat{\alpha}_b^{\mathcal{C}}$ when $\mathcal{C} =$ `UN`, `EqOpt`, `DP` under different sets of parameters. Specifically, in Table 5.1, $n_a = n_b$, $u_+ = u_-$, $[\mu_s^0, \mu_s^1, \sigma_s] = [-5, 5, 5]$, $\forall s \in \{a, b\}$ and transitions satisfying either Condition 1a) or 1b) are randomly generated; in Table 5.2, transitions satisfying Condition 1b) and $f_s^y(x)$ that satisfy Condition 2 are randomly generated, $\frac{u_+}{u_-}$ also satisfies the condition in Theorem 21. These results are consistent with Theorem 20 and 21.

Table 5.1: $\widehat{\alpha}_a^{\mathcal{C}} - \widehat{\alpha}_b^{\mathcal{C}}$ when $\mathcal{C} = \texttt{UN}, \texttt{EqOpt}, \texttt{DP}$: $f_a^y = f_b^y$ and $T_a^{yd} \neq T_b^{yd}$.

					Condition 1a)				
UN ($\times 10^{-2}$)	-18.45	16.89	19.82	-7.21	-16.34	-26.56	16.66	-6.03	-38.63
EqOpt ($\times 10^{-2}$)	-21.11	19.13	21.78	-7.62	-18.56	-29.21	18.14	-6.28	-41.52
DP ($\times 10^{-2}$)	-27.98	23.11	25.65	-8.90	-23.11	-33.22	21.09	-6.66	-43.35
					Condition 1b)				
UN ($\times 10^{-2}$)	-19.05	18.18	-0.70	-58.80	-40.91	61.30	12.82	-44.67	2.66
EqOpt ($\times 10^{-2}$)	-18.40	17.98	-0.64	-57.62	-34.50	48.66	12.35	-41.43	2.61
DP ($\times 10^{-2}$)	-17.52	17.73	-0.57	-55.62	-28.97	36.10	11.69	-37.97	2.57

Table 5.2: $\widehat{\alpha}_a^{\mathcal{C}} - \widehat{\alpha}_b^{\mathcal{C}}$ when $\mathcal{C} = \texttt{UN}, \texttt{EqOpt}, \texttt{DP}$: $f_a^y \neq f_b^y$ and $T_a^{yd} = T_b^{yd}$ under Condition 1b).

UN ($\times 10^{-2}$)	1.88	26.35	2.12	0.38	5.64	12.35	11.70	0.20	4.12
EqOpt ($\times 10^{-2}$)	0.57	17.43	1.75	0.32	5.05	7.81	7.21	0.18	1.68
DP ($\times 10^{-4}$)	16.26	18.29	-5.94	-0.93	-2.25	1.47	0.92	-1.68	-0.80

FICO scores data. We use the FICO score dataset [122] to study the long-term impact of fairness constraints EqOpt and DP and other interventions on loan repayment rates in the Caucasian group $\mathcal{G}_C$ and the African American group $\mathcal{G}_{AA}$. FICO scores are widely used in the US to assess an individual's creditworthiness. With the pre-processed data in [57], we simulate a dataset with loan repayment records and credits scores. We first compute group proportions $n_C = 0.88, n_{AA} = 0.12$, the initial qualification (loan repayment) rates $\alpha_C(0) = 0.76, \alpha_{AA}(0) = 0.34$ and estimate the feature distributions $f_s^y(x)$ with beta distributions based on the simulated data, as shown in Figure 5.3. Then, we compute the optimal UN, EqOpt, DP threshold according to Eqn. (5.3). Consequently, with the dynamics (5.4), we update the qualification rates in both groups. This process proceeds and qualification rates in both groups change over time.

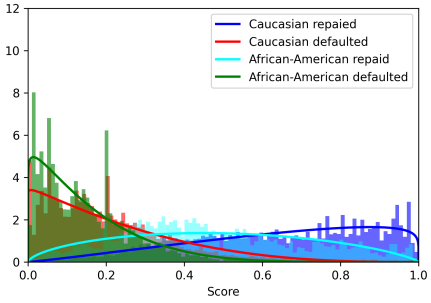
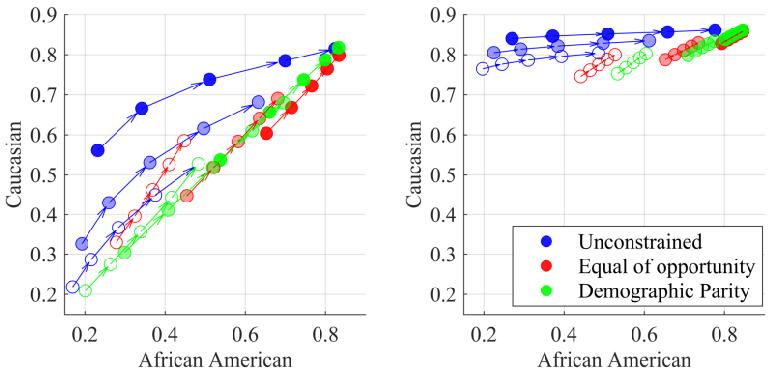


Figure 5.3: The feature distributions: the scores are rescaled so that they are between 0 and 1.

Our results show consistent findings with studies in loan repayment literature [52, 119]. Specifically, [119] studied the loan repayment in group lending and pointed out that in practice effective training and leadership among the groups who were issued loans can increase their willingness to pay and improve the group repayment rate. Similar conclusion is also suggested by [52]. In our model, these interventions can be regarded as stimulating transitions (i.e., T_s^{y1}) to improve the future repayment rates. And the scenarios under such intervention would satisfy Condition 1b). Figure 6.10 illustrates the equilibria $(\widehat{\alpha}_{AA}, \widehat{\alpha}_C)$ under different sets of transitions (including demographic-invariant (*D-invariant*) and demographic-variant (*D-variant*) transitions). Their specific values are listed below, where the system has an equilibrium in all cases.

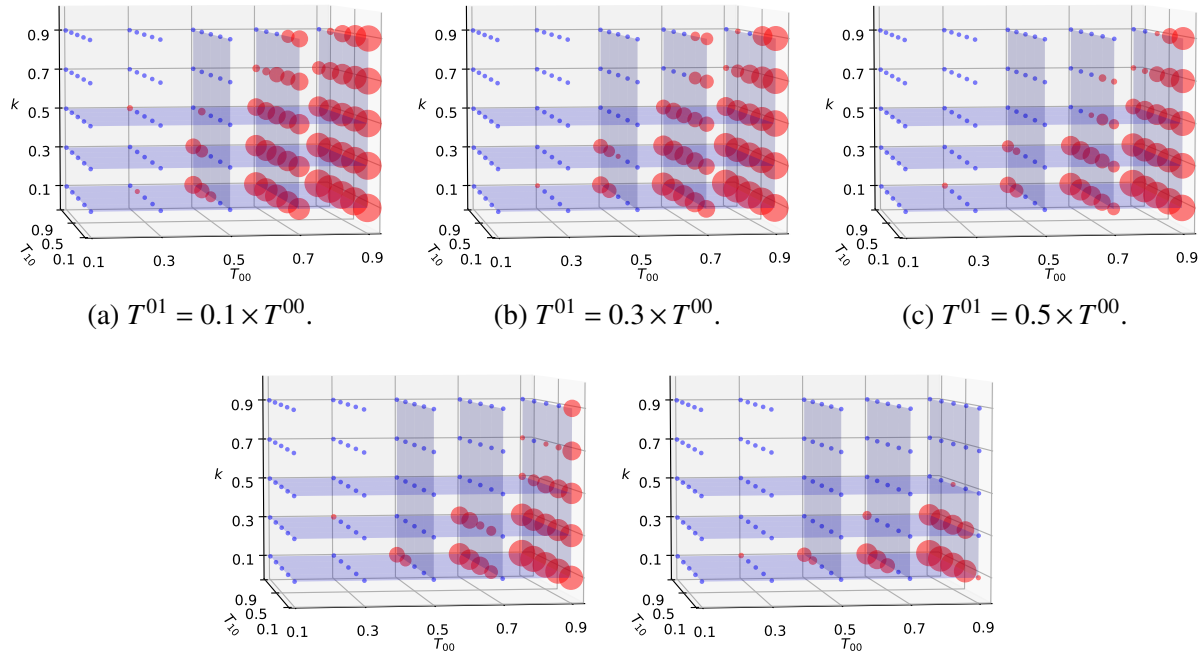


(a) D-invariant transitions (b) D-variant transitions

Figure 5.4: Results on the FICO dataset: Points are the equilibria of repayment rates in $\mathcal{G}_{AA}, \mathcal{G}_C$ under Condition 1b) with different transitions. Arrows indicate the direction of increasing T_s^{01}; a more transparent point represents the smaller value of T_s^{10}. In panel a, $T_{AA}^{yd} = T_C^{yd}$, while in panel b, $T_{AA}^{yd} < T_C^{yd}$.

$$
\begin{aligned}
\text{D-invariant:} \quad & T_s^{00} = 0.1, T_s^{11} = 0.9, \quad T_s^{10}, T_s^{01} \in \{0.1, 0.5, 0.9\}, s \in \{AA, C\} \\
\text{D-variant:} \quad & T_{AA}^{00} = 0.1, T_{AA}^{11} = 0.9, \quad T_{AA}^{10}, T_{AA}^{01} \in \{0.20, 0.53, 0.85\} \\
& T_C^{00} = 0.4, T_C^{11} = 0.9, \quad T_C^{10}, T_C^{01} \in \{0.45, 0.65, 0.85\}
\end{aligned}
$$

It shows that under Condition 1b), increasing the transition T_s^{01} always increases qualification rates, and `DP` in general can result in a more equitable equilibrium than `EqOpt`. Figure 5.4a shows that in Demographic-invariant (D-invariant) transition cases ($T_{AA}^{yd} = T_C^{yd}$): (1) $\mathcal{G}_{AA}$ always remains as disadvantaged group; (2) when T_s^{10} is small, the inequality under `UN` optimal policies is small and the intervention on T_s^{01} only has minor effects on equality; when T_s^{10} is large (darker blue points), varying T_s^{01} can affect disparity significantly; (3) imposing `DP` attains equitable equilibria in general, which is robust to transitions and consistent with the conclusion in [113]; (4) when T_s^{10} is small, imposing `EqOpt` exacerbates inequality as T_s^{01} increases; while T_s^{10} is sufficient large, equality



(a) $T^{01} = 0.1 \times T^{00}$. (b) $T^{01} = 0.3 \times T^{00}$. (c) $T^{01} = 0.5 \times T^{00}$.

(d) $T^{01} = 0.7 \times T^{00}$. (e) $T^{01} = 0.9 \times T^{00}$.

Figure 5.5: The oscillation level of recidivism rates under different transitions. In each panel, scalar k denotes the ratio, of which $T^{11} = k \times T^{10}$.

can be attained and robust to transitions. In Figure 5.4b, it shows that in D-variant transition cases, by setting $T^{yd}_{AA} < T^{yd}_{C}$, the inequality between $\mathcal{G}_{AA}$ and $\mathcal{G}_{C}$ further gets reinforced. In summary, the effectiveness of such intervention (increasing T^{01}_{s}) on promoting equality highly depends on the value of T^{10}_{s} and policies.

The COMPAS data. Our second set of experiments is conducted on a multivariate recidivism prediction dataset from Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) [7]. We again use this static and high-dimensional dataset to create a simulated decision-making process as the FICO experiments.

Specifically, from the raw data we calculate the initial qualification (recidivism) rate and train optimal classifier using a logistic regression model, based on which recidivism rate is updated according to Eqn. (5.4) under a given set of transitions. In the context of recidivism prediction, we consider all the possible types of transitions under an unconstrained policy, i.e., transitions satisfying

conditions 1a)-d). The classifier decision here corresponds to incarceration based on predicted likelihood of recidivism: the higher the predicted recidivism, the more likely an incarceration decision. In subsequent time steps, the data is re-sampled from the raw data proportional to the updated recidivism rates. This process repeats and the group recidivism rates change over time.

Table 5.3: osi/osi$_H$/osi$_L$ is the percentage that oscillation occurs among 125 set of different transitions under policy UN/UN$_{\theta_H}$/UN$_{\theta_L}$. Among transitions that lead to stable equilibrium, Column 2/Column 3 shows the percentage that UN$_{\theta_H}$/ UN$_{\theta_L}$ results in lower recidivism compared with UN.

	$\widehat{\alpha}_{\theta_H} < \widehat{\alpha}$	$\widehat{\alpha}_{\theta_L} < \widehat{\alpha}$	osi	osi$_H$	osi$_L$
1a)	0	1	0.29	0.12	0.36
1b)	0.99	0.01	0	0	0
1c)	0.37	0.28	0	0	0
1d)	0.79	0.63	0.06	0	0.13

Our results here primarily serve to highlight the complexity in such a decision making system. In particular, we see that an equilibrium may not exist and under some transitions the qualification rate may oscillate. Specifically, Table 5.3 shows that Proposition 1 holds under Condition 1a)-b); there is no oscillation under Condition 1b)- c); under Condition 1c)-d), there is more uncertainty.

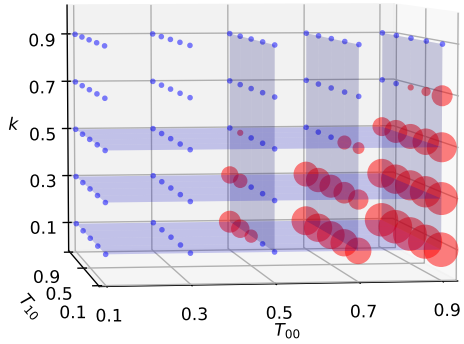


Figure 5.6: $T^{y1} = k \times T^{y0}$, $y = 0, 1$. The oscillation level of recidivism rates in the long run is represented by the size of red circles, the bigger size means the severer oscillation. The blue dots indicate the cases with a unique equilibrium.

To further explore when the system is in an equilibrium state under unconstrained optimal policy, we consider a set of transitions with T^{00} and T^{10} taking the values 0.1, 0.3, 0.5, 0.7 and 0.9. Figure 5.6 shows the results when $T^{01} = k \times T^{00}$ and $T^{11} = k \times T^{10}$, $k \in \{0, 1, 0.3, 0.5, 0.7, 0.9\}$. We find that when Corollary 2 is satisfied, e.g., when $k \geq 0.5$, most of the systems have a unique equilibrium (blue dot). Moreover, when $T^{00} \leq 0.5$, the system is also mostly in the unique equilibrium state. For other transitions, the system oscillates between two states (red circle). We also show the results under other combinations of T^{01} and T^{11} in Figure 5.5.

Next, we study the impact of policy interventions in cases with equilibrium. We randomly choose the transitions under which the system has an equilibrium and then apply the unconstrained policy with optimal threshold (classifier threshold 0.5), a higher and a lower threshold (classifier thresholds 0.8 and 0.2 respectively) compared to the optimum respectively. The results are show in Table 5.4, where UN indicates the unconstrained policy with the optimal threshold, $\mathtt{UN}_{\theta_H}$ means the policy with a higher threshold, and $\mathtt{UN}_{\theta_L}$ the policy with a lower threshold.

Table 5.4: Recidivism rates in the long run under different policies of 5 independent runs of experiments.

	$\mathtt{UN}^*$	$\mathtt{UN}_{\theta_H}$	$\mathtt{UN}_{\theta_L}$
$\widehat{\alpha}_1$	0.164	0.166	0.147
$\widehat{\alpha}_2$	0.343	0.356	0.307
$\widehat{\alpha}_3$	0.230	0.246	0.162
$\widehat{\alpha}_4$	0.306	0.3415	0.156
$\widehat{\alpha}_5$	0.162	0.166	0.140

CHAPTER 6

Impact of Fairness Interventions on Strategic Manipulation

6.1 Introduction

In Chapters 4 and 5, we studied two types of interactions between individuals and ML system, where ML decisions have *downstream* impacts, i.e., impact on participation (Chapter 4) or impact on qualification (Chapter 5), on individuals' behaviors. Such impacts are then captured in the dataset used for building the future ML systems. In practice, when ML systems are deployed to make decisions about people, there is a requirement for transparency in terms of how decisions are reached given input. As a result, given (partial) information about an algorithm, individuals subject to its decisions can and will adapt their behavior by strategically manipulating their data in order to obtain favorable decisions [19,21,25,32,53,56,66,96,110,111]. This strategic behavior in turn hurts the performance of ML models and diminishes their utility. Such a phenomenon has been widely observed in real-world applications, and is known as *Goodhart's law*, which states "once a measure becomes a target, it ceases to be a good measure" [127]. For instance, a hiring or admissions practice that heavily depends on GPA might motivate students to cheat on exams; not accounting for such manipulation may result in disproportionate hiring of under-qualified individuals. A strategic decision maker is one who anticipates such behavior and thus aims to make its ML models robust to such strategic manipulation.

In this chapter, we focus on the design of (fair) machine learning models in the presence of strategic manipulation. Same as Chapters 4 and 5, we consider a decision maker whose goal is to select qualified individuals based on a given set of features. Given knowledge of the selection policy, individuals can tailor their behavior and manipulate their features to receive favorable decisions.

We shall assume that this feature manipulation does not affect an individual's true qualification state. We say the decision maker (and its policy) is *strategic* if it anticipates such manipulation; it is *non-strategic* if it does not take into account individuals' manipulation in its policies.

We adopt a typical two-stage (Stackelberg) game setting where the decision maker commits to its policies, following which individuals best-respond. Under this model, we study the impact of fairness intervention on different social groups in the presence of strategic manipulative behavior, and explore the role of fairness intervention in (dis)incentivizing such manipulation. We aim to answer the following questions: how does the anticipation of individuals' strategic behavior impact a decision maker's utility, and the resulting policies' fairness properties? How is the Stackelberg equilibrium affected when fairness constraints are imposed? Can fairness intervention serve as incentives/disincentives for individuals' strategic manipulation?

Our main contributions and findings are as follows.

1. We formulate a Stackelberg game to model the interaction between a decision maker and strategic individuals (Section 6.3). We characterize both strategic (fair) optimal policies and non-strategic (fair) optimal policies of the decision maker, and individuals' best response (Section 6.4, Lemmas 8-11).
2. We study the impact of the decision maker's anticipation of individuals' strategic manipulation by comparing non-strategic with strategic policies (Section 6.5):
    - We show that compared to the non-strategic policy, the strategic policy always disincentivizes manipulative behavior, but that it over (resp. under) selects when a population is majority-qualified (resp. majority-unqualified)[1] (Theorem 22).
    - We show that the anticipation of manipulation can adversely affect the fairness of a strategic policy: when one group is majority-qualified while the other is majority-unqualified, we identify conditions under which strategic policy always *worsens unfairness* (Theorem 23); on the other hand, when both groups are majority-unqualified, we show that it is possible to use the strategic policy to *mitigate unfairness* and even flip the disadvantaged group (Theorem 24).
3. We study the impact of fairness interventions on policies and individuals' manipulation (Section 6.6).
    - If a decision maker lacks information or awareness to anticipate strategic behavior (but which in fact exists), we identify conditions under which such non-strategic decision

[1] A group is majority-(un)qualified if the majority of that population is (un)qualified.

maker benefits from using fairness constrained policies rather than unconstrained policies (Theorem 25).

- By comparing individuals' responses to a strategic policy with and without fairness intervention, we show that fairness interventions can serve as (dis)incentives for manipulation, and identify scenarios under which a strategic fair policy can (dis)incentivize manipulation compared to a strategic policy (Theorems 26 and 27).

4. We examine our theoretical findings using both synthetic and real-world data (Section 6.7).

The remainder of this chapter is organized as follows. Section 6.2 presents related work. Section 6.3 formulates the problem. Section 6.4 presents four types of (non-)strategic (fair) policies. The impact of decision maker's anticipation of manipulative behavior is analyzed and presented in Section 6.5. The impact of fairness interventions on policies and individuals' manipulation is studied in Section 6.6. Experiments are presented in Section 6.7. All proofs can be found in Appendix E.

6.2 Related Work

Our work closely connects to the literature in classification problems in the presence of strategic manipulation. [56] formulates such problem as a Stackelberg competition between the decision maker and individuals, where the decision maker publishes the classifier first, and individuals after observing the classifier can manipulate their features at costs to maximize their utilities. Different from the Stackelberg formulation in our work, manipulation cost in [56] is modeled as a deterministic function of change in features before and after manipulation. The decision maker aims to find an optimal classifier such that the classification accuracy is maximized when individuals best respond, and the learning algorithms are developed in [56]. [32] extends this strategic classification to an online setting, where data arrives sequentially and only the manipulated data is revealed. An online convex classification learning algorithm is designed such that the averaged regret diminishes in the long run. [66, 111] extend [56] by assuming individuals from the different social groups have different costs in manipulation, and the disparate impacts on different groups are studied. [19] explores the role of randomness in strategic classification and focuses on randomized classifiers. It shows that randomness can improve classification accuracy and mitigate the disparate effects incurred by manipulation costs across different groups in strategic settings.

Note that the manipulation does not affect an individual's underlying label in the works mentioned above, i.e., strategic manipulation is viewed as gaming. In contrast, another line of research [27, 53, 96] considers a setting where the individual's label (qualification) changes in accordance with the strategic behavior. Specifically, the goal of the decision maker is to design a classifier such that individuals are incentivized to behave toward directions that improve the underlying qualifications [53, 96]. [25, 110] consider both types of strategic behavior: gaming without changing labels and improvement. Specifically, [25] trains classifiers that disincentivize the manipulation while incentivizing improvement. [110] proposes a causal framework for distinguishing between gaming and improvement.

In Stackelberg game formulations, the decision maker always moves first and individuals respond after decision maker's action has been disclosed. Instead, [20, 27, 104] consider scenarios where both individuals and the decision maker act simultaneously. They formulate the strategic interaction between individuals and decision maker as a game and study the Nash equilibria of the game. [27] considers a setting where individuals are from two social groups which are identical in nature but one group suffers from the negative stereotype. It shows that such stereotype results in different equilibria of two groups. The impact of *demographic parity* fairness is also examined in [27]. [104] studies a similar game, but it assumes two groups can be different in feature distributions and manipulation costs.

6.3 Problem Formulation

Consider two demographic groups $\mathcal{G}_a$, $\mathcal{G}_b$ distinguished by a sensitive attribute $S \in \{a, b\}$ (e.g., gender), with fractions $n_s = \Pr(S = s)$ of the population. An individual from either group has observable features $X \in \mathbb{R}^d$ and a hidden qualification state $Y \in \{0, 1\}$. Let $\alpha_s = P_{Y|S}(1|s)$ be the qualification rate of $\mathcal{G}_s$, and $f_s^y(x) = P_{X|Y,S}(x|y, s)$. A decision maker makes a decision $D \in \{0, 1\}$ ("0" being negative/reject and "1" positive/accept) for an individual using a group-dependent policy $\pi_s(x) = P_{D|X,S}(1|x, s)$. An individual's action is denoted by $M \in \{0, 1\}$, with $M = 1$ indicating manipulation and $M = 0$ otherwise. Note that in our context manipulation does not change the true qualification state Y.

Best response. An individual in $\mathcal{G}_s$ incurs a random cost $C_s \geq 0$ when manipulating its features, with probability density function (PDF) $P_{C_s}(c)$ and cumulative density function (CDF) $\mathbb{F}_{C_s}(c) = \int_0^c P_{C_s}(z)dz$. The realization of this random cost is known to an individual when determining its

action M; the decision maker on the other hand only knows the overall cost distribution of each group. Thus the response that the decision maker anticipates (from the group as a whole or from a randomly selected individual) is expressed as follows, whereby given policy π_s, an individual in $\mathcal{G}_s$ will manipulate its features if doing so increases its utility:

$$wP_{D|Y,M,S}(1|y,1,s) - C_s \geq wP_{D|Y,M,S}(1|y,0,s).$$

Here $w > 0$ is a fixed benefit to the individual associated with a positive decision $D = 1$ (the benefit is 0 otherwise); without loss of generality we will let $w = 1$. In other words, the best response the decision maker expects from the individuals of $\mathcal{G}_s$ with qualification y is their probability of manipulation, denoted by p_s^y and written as:

$$p_s^y(\pi_s) = \Pr\left(C_s \leq P_{D|Y,M,S}(1|y,1,s) - P_{D|Y,M,S}(1|y,0,s)\right).$$

We will assume that only those unqualified may choose to manipulate, and they do so by imitating the features of the qualified, i.e., $P_{M|Y,S}(1|1,s) = 0$, $P_{X|Y,M,S}(x|0,0,s) = P_{X|Y,S}(x|0,s)$ and $P_{X|Y,M,S}(x|0,1,s) = P_{X|Y,S}(x|1,s)$. This would mean, for instance, that those qualified have no incentive to cheat on an exam, whereas those unqualified may choose to cheat by copying answers from the qualified. This assumption is inspired by the *imitative learning* behavior observed in social learning, whereby new behaviors are acquired by copying social models' actions [49, 50].

Importantly, the feature distributions of unqualified individuals are different before and after manipulation. To avoid confusion, we will always use $f_s^y(x) = P_{X|Y,S}(x|y,s)$ to denote the conditional probability distributions of features *before* manipulation. The feature distribution of those unqualified after manipulation becomes $(1 - p_s^0(\pi_s))f_s^0(x) + p_s^0(\pi_s)f_s^1(x)$.

Optimal (fair) policy. The decision maker receives a true-positive (resp. false-positive) benefit (resp. penalty) u_+ (resp. u_-) when accepting a qualified (resp. unqualified) individual. Its utility, denoted by $R(D,Y)$, is $R(1,1) = u_+$, $R(1,0) = u_-$, $R(0,0) = R(0,1) = 0$. The decision maker aims to find optimal policies for the two groups such that its expected total utility $\mathbb{E}[R(D,Y)]$ is maximized.

As mentioned earlier, there are two types of decision makers, strategic and non-strategic: A *strategic decision maker* anticipates strategic manipulation, has perfect information on the manipulation cost distribution, and accounts for this in determining policies, while a *non-strategic decision maker* ignores manipulative behavior in determining its policies. Either type may further

impose a fairness constraint $\mathcal{C}$, to ensure that π_a and π_b satisfy the following:

$$\mathbb{E}_{X\sim\mathcal{P}_a^{\mathcal{C}}}[\pi_a(X)] = \mathbb{E}_{X\sim\mathcal{P}_b^{\mathcal{C}}}[\pi_b(X)]\ , \tag{6.1}$$

where $\mathcal{P}_s^{\mathcal{C}}$ is some probability distribution over X in accordance with the fairness constraint $\mathcal{C}$. Many common fairness notions can be written in this form, e.g., equal opportunity (`EqOpt`) [57] where $\mathcal{P}_s^{\texttt{EqOpt}}(x) = P_{X|Y,S}(x|1,s)$, or demographic parity (`DP`) [11] where $\mathcal{P}_s^{\texttt{DP}}(x) = P_{X|S}(x|s)$.

The above leads to four types of optimal policies a decision maker can use, which we consider in this chapter: (1) a non-strategic policy; (2) a non-strategic fair policy; (3) a strategic policy; (4) a strategic fair policy. These are detailed in Section 6.4.

The Stackelberg game. The interaction between the decision maker and the individuals consists of the following two stages in sequence: (i) The former publishes its policies (π_a, π_b), which may be strategic or non-strategic, and may or may not satisfy a fairness constraint, and (ii) the latter, while observing the published policies and their realized costs, decide whether to manipulate their features.

6.4 The Four Types of (Non-)Strategic (Fair) Policies

Non-strategic policy. A decision maker who does not account for individuals' strategic manipulation optimizes the following expected utility over $\mathcal{G}_s$

$$\widehat{U}_s(\pi_s) = \int_X \Big[u_+\alpha_s f_s^1(x) - u_-(1-\alpha_s) f_s^0(x)\Big]\pi_s(x)dx.$$

Define $\mathcal{G}_s$'s *qualification profile* as $\gamma_s(x) = P_{Y|X,S}(1|x,s)$. Based on Chapter 5, we can show that the non-strategic policy $\widehat{\pi}_s^{\texttt{UN}} = \text{argmax}_{\pi_s}\widehat{U}_s(\pi_s)$ is in the form of a threshold policy, i.e., $\widehat{\pi}_s^{\texttt{UN}}(x) = \mathbf{1}(\gamma_s(x) \geq \frac{u_-}{u_+ + u_-})$. Throughout the chapter, we will present results in the one dimensional feature space. The same can be generalized to high dimensional spaces (Appendix E.1).

Assumption 11. *$f_s^1(x)$, $f_s^0(x)$ are continuous and satisfy the strict monotone likelihood ratio property, i.e., $\frac{f_s^1(x)}{f_s^0(x)}$ is increasing in $x \in \mathbb{R}$. Let unique x_s^* be s.t. $\frac{f_s^1(x_s^*)}{f_s^0(x_s^*)} = 1$.*

Assumption 11 is relatively mild and can be satisfied by distributions such as exponential and Gaussian, and has been widely used [10, 27, 75, 88, 155]. It implies that an individual is more likely

to be qualified as their feature value increases. Under Assumption 11, the threshold policy can be written as $\pi_s(x) = \mathbf{1}(x \geq \theta_s)$ for some $\theta_s \in \mathbb{R}$. Throughout the chapter, we assume Assumption 11 holds and focus on threshold policies. We will frequently use θ_s to denote policy π_s.

Under Assumption 11, the thresholds for non-strategic policies are characterized as follows.

Lemma 8. *Let $(\widehat{\theta}_a^{UN}, \widehat{\theta}_b^{UN})$ be non-strategic optimal thresholds. Then $\frac{f_s^1(\widehat{\theta}_s^{UN})}{f_s^0(\widehat{\theta}_s^{UN})} = \frac{u_-(1-\alpha_s)}{u_+\alpha_s}$.*

Non-strategic fair policy. Denoted as $(\widehat{\pi}_a^{\mathcal{C}}, \widehat{\pi}_b^{\mathcal{C}})$, this is found by maximizing the total utility subject to fairness constraint $\mathcal{C}$, i.e., $(\widehat{\pi}_a^{\mathcal{C}}, \widehat{\pi}_b^{\mathcal{C}}) = \text{argmax}_{(\pi_a,\pi_b)} n_a \widehat{U}_a(\pi_a) + n_b \widehat{U}_b(\pi_b)$ such that Eqn (6.1) holds. Based on Chapter 5, It can be shown that for `EqOpt` and `DP` fairness, the optimal fair policies are also threshold policies and can be characterized by the following.

Lemma 9 (Lemma 7, Chapter 5)**.** *Let $(\widehat{\theta}_a^{\mathcal{C}}, \widehat{\theta}_b^{\mathcal{C}})$ be thresholds in non-strategic optimal fair policies. These satisfy*

$$\sum_{s=a,b} n_s \left(\frac{u_+\alpha_s f_s^1(\widehat{\theta}_s^{\mathcal{C}}) - u_-(1-\alpha_s) f_s^0(\widehat{\theta}_s^{\mathcal{C}})}{\mathcal{P}_s^{\mathcal{C}}(\widehat{\theta}_s^{\mathcal{C}})} \right) = 0 \, .$$

Strategic policy. Let $p_s^0 := P_{M|Y,S}(1|0,s)$, the probability that unqualified individuals in $\mathcal{G}_s$ manipulate. The decision maker's expected utility over $\mathcal{G}_s$ under $\pi(x) = \mathbf{1}(x \geq \theta)$ is as follows:

$$\begin{aligned} U_s(\theta) &= u_+\alpha_s(1 - \mathbb{F}_s^1(\theta)) - u_-(1-\alpha_s)\Big(1 - \mathbb{F}_s^0(\theta)(1-p_s^0) - \mathbb{F}_s^1(\theta)p_s^0\Big) \\ &= \widehat{U}_s(\theta) - u_-(1-\alpha_s)\Big(\mathbb{F}_s^0(\theta) - \mathbb{F}_s^1(\theta)\Big)p_s^0 \end{aligned}$$

where $\mathbb{F}_s^y(x) = \int_{-\infty}^{x} f_s^y(z)dz$ denotes the CDF.

Define the *manipulation benefit* as $\Delta_s(\theta) := \mathbb{F}_s^0(\theta) - \mathbb{F}_s^1(\theta)$; this represents the additional benefit an individual gains from manipulation. The unqualified individuals' best-response to a policy with threshold θ will be $p_s^0(\theta) = \mathbb{F}_{C_s}\Big(P_{D|Y,M,S}(1|0,1,s) - P_{D|Y,M,S}(1|0,0,s)\Big) = \mathbb{F}_{C_s}(\Delta_s(\theta))$. This manipulation probability $p_s^0(\theta)$ is single-peaked with maximum occurring at x_s^*, and $\lim_{\theta\to-\infty} p_s^0(\theta) = \lim_{\theta\to+\infty} p_s^0(\theta) = 0$, meaning that when the threshold is sufficiently low or high, unqualified individuals are less likely to manipulate their features. Plugging this in the decision maker's utility, we have

$$U_s(\theta) = \widehat{U}_s(\theta) - \underbrace{u_-(1-\alpha_s)\Delta_s(\theta)\mathbb{F}_{C_s}(\Delta_s(\theta))}_{\textbf{term 2}:=\Psi_s(\Delta_s(\theta))} \, . \tag{6.2}$$

Define a function $\Psi_s(z) := u_-(1-\alpha_s)\mathbb{F}_{C_s}(z)z$, then **term 2** in Eqn. (6.2) can be written as $\Psi_s(\Delta_s(\theta))$, and can be interpreted as the additional loss incurred by the decision maker due to manipulation (equivalently, the average manipulation gain by group $\mathcal{G}_s$). Further, consider its first order derivative $\Psi'_s(z) = u_-(1-\alpha_s)\frac{d\mathbb{F}_{C_s}(z)z}{dz} = u_-(1-\alpha_s)(\mathbb{F}_{C_s}(z) + zP_{C_s}(z))$. This $\Psi'_s(\Delta_s(\theta))$ indicates the decision maker's *marginal loss* caused by strategic manipulation (equivalently, the *marginal* manipulation gain of $\mathcal{G}_s$). The thresholds for strategic policies are characterized as follows.

Lemma 10. *For $(\theta_a^{UN}, \theta_b^{UN})$, the strategic optimal thresholds, $\frac{f_s^1(\theta_s^{UN})}{f_s^0(\theta_s^{UN})} = \frac{u_-(1-\alpha_s)-\Psi'_s(\Delta_s(\theta_s^{UN}))}{u_+\alpha_s-\Psi'_s(\Delta_s(\theta_s^{UN}))}$.*

Strategic fair policy. The strategic fair thresholds $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$ are found by maximizing the total expected utility subject to fairness constraint $\mathcal{C}$, i.e., $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}}) = \text{argmax}_{(\theta_a,\theta_b)} n_a U_a(\theta_a) + n_b U_b(\theta_b)$ such that Eqn. (6.1) holds. They can be characterized by the following.

Lemma 11. *Let $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$ be thresholds in strategic optimal fair policies. These satisfy*

$$\sum_{s=a,b} n_s \left(\frac{f_s^0(\theta_s^{\mathcal{C}}) - f_s^1(\theta_s^{\mathcal{C}})}{\mathcal{P}_s^{\mathcal{C}}(\theta_s^{\mathcal{C}})} \Psi'_s(\Delta_s(\theta_s^{\mathcal{C}})) + \frac{u_+\alpha_s f_s^1(\theta_s^{\mathcal{C}}) - u_-(1-\alpha_s) f_s^0(\theta_s^{\mathcal{C}})}{\mathcal{P}_s^{\mathcal{C}}(\theta_s^{\mathcal{C}})} \right) = 0.$$

Note that in addition to $(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ and $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$, the equations in Lemmas 10 and 11 may be satisfied by other threshold pairs that are not optimal. We discuss this further in the next section.

6.5 Impact of the Decision Maker's Anticipation of Manipulative Behavior

Impact on the optimal policy & utility function. We first compare strategic policy θ_s^{UN} to non-strategic policy $\widehat{\theta}_s^{\text{UN}}$, and examine how the policy and the decision maker's expected utility differ.

Assumption 12. $\Psi'_s(z) < \infty$ *is non-decreasing over* $[0, \max_\theta \Delta_s(\theta)]$.

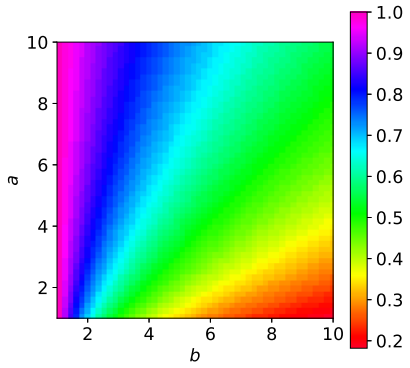
For any threshold θ, $\Delta_s(\theta)$ represents the manipulation benefit of $\mathcal{G}_s$; those in $\mathcal{G}_s$ choose to manipulate if $C_s \le \Delta_s(\theta)$. Therefore, $\max_\theta \Delta_s(\theta)$ indicates the maximum additional benefit an individual in $\mathcal{G}_s$ may gain from manipulation. As $\Psi'_s(\Delta_s(\theta))$ represents the marginal manipulation gain of $\mathcal{G}_s$ on average, Assumption 12 means that a group's *marginal* manipulation gain does

not decrease as manipulation benefit increases. Examples (e.g., beta/uniformly distributed cost) satisfying this assumption can be found in the following.[2]

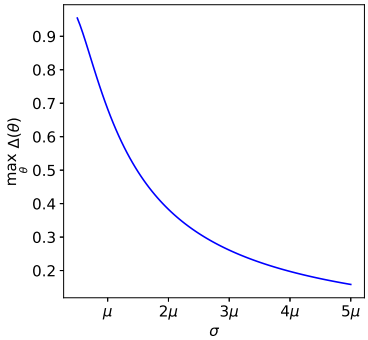
Remark 1. *For simplicity, we drop subscript s in the following.*

Example 1: *cost* $C \sim U[0,\overline{c}]$. *In this case,* $\Psi'(z) = u_-(1-\alpha)\frac{2}{\overline{c}}z$ *is non-decreasing.*

Example 2: *cost* $C \sim Beta(a,b)$ *with* $a \in [1,10]$, $b \in [1,10]$.



(a) $\overline{\Delta}$ that ensures $\Psi'(z)$ to be non-decreasing over $[0,\overline{\Delta}]$ when $C \sim \text{Beta}(a,b)$, $a \in [1,10], b \in [1,10]$



(b) $\max_\theta \Delta(\theta)$ for Gaussian distributed feature where $X|Y=1 \sim \mathcal{N}(\mu,\sigma^2)$, $X|Y=0 \sim \mathcal{N}(-\mu,\sigma^2)$, and $\mu > 0$.

For Beta distributed cost and Gaussian distributed features, above figures show that Assumption 12 is relatively mild. For example, when $C \sim Beta(8,3)$, *the left plot shows that* $\Psi'(z)$ *is non-decreasing over* $[0,0.82]$. *For features that follow Gaussian distributions* $X|Y=1 \sim \mathcal{N}(\mu,\sigma^2)$ *and* $X|Y=0 \sim \mathcal{N}(-\mu,\sigma^2)$, *the condition is satisfied as long as* $\sigma > 0.72\mu$.

Other examples: *There are many other probability density distributions with support* $[0,1]$ *or* $[0,\infty)$ *that could satisfy this condition, such as beta prime distribution, gamma distribution, chi distribution, chi-squared distribution, etc.*

Note that under Assumption 12, $\Psi'_s(0) = 0$ and $\Psi'_s(\Delta_s(\theta))$ is single-peaked with maximum occurring at x^*_s. We assume it holds in Sections 6.5 and 6.6.

Theorem 22. *Let* $\overline{\Psi'_s} = \max_\theta \Psi'_s(\Delta_s(\theta))$, *and* $\delta_u = \frac{u_-}{u_-+u_+}$.

1. If $\alpha_s = \delta_u$, *then* $\theta^{UN}_s = \widehat{\theta}^{UN}_s = x^*_s$ *when* $\overline{\Psi'_s} \le u_-(1-\alpha_s)$, *and* $\theta^{UN}_s \in \{\underline{x_s}, \overline{x_s}\}$ *otherwise.*

[2] In economics, a choice of *generalized beta distribution* is common to model costs (e.g., healthcare costs [72]). In addition to uniformly distributed C_s (same as [104]), we consider beta distributed C_s in our experiments.

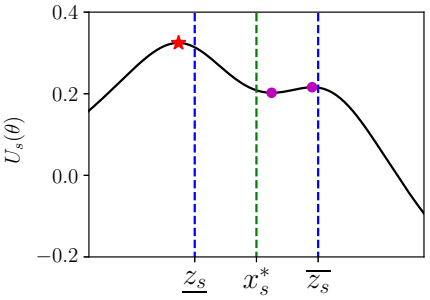
2. *If $\alpha_s < \delta_u$, then $\theta_s^{UN} > \widehat{\theta}_s^{UN} > x_s^*$. Moreover, if $\overline{\Psi'_s} > u_-(1-\alpha_s)$, then $\theta_s^{UN} > \overline{x_s}$ and $U_s(\theta)$ may have additional extreme points in $(\underline{x_s}, x_s^*)$; otherwise $\widehat{\theta}_s^{UN}$ is the unique extreme point of $U_s(\theta)$.*

3. *If $\alpha_s > \delta_u$, then $\theta_s^{UN} < \widehat{\theta}_s^{UN} < x_s^*$. Moreover, if $\overline{\Psi'_s} > u_+\alpha_s$, then $\theta_s^{UN} < \underline{z_s}$ and $U_s(\theta)$ may have additional extreme points in $(x_s^*, \overline{z_s})$; otherwise $\widehat{\theta}_s^{UN}$ is the unique extreme point of $U_s(\theta)$.*

Here $\underline{x_s} < \overline{x_s}, \underline{z_s} < \overline{z_s}$ are defined such that $\Psi'_s(\Delta_s(\underline{x_s})) = \Psi'_s(\Delta_s(\overline{x_s})) = u_-(1-\alpha_s)$, $\Psi'_s(\Delta_s(\underline{z_s})) = \Psi'_s(\Delta_s(\overline{z_s})) = u_+\alpha_s$.

We note that even though $\widehat{U}_s(\theta)$ (non-strategic utility) and $\Psi_s(\Delta_s(\theta))$ are single-peaked and have unique extreme points, their difference $U_s(\theta)$ (Eqn.(6.2)) may have multiple extreme points. As we will see later, this results in strategic and non-strategic policies having different properties in many aspects.

An example of $U_s(\theta)$ is shown to the right, where $f_s^1(x), f_s^0(x)$ are Gaussian distributed with the same variance 4.7^2 and means $5, -5$ respectively. $C_s \sim \text{Beta}(10,4)$, $\alpha_s = 0.6$ and $u_- = u_+$. The red star is the optimal threshold $\theta_s^{\text{UN}} < \underline{z_s}$; two magenta dots are other extreme points of $U_s(\theta)$, which are in $(x_s^*, \overline{z_s})$. Theorem 22 states that $U_s(\theta)$ has multiple extreme points if $\overline{\Psi'_s}$ is sufficiently large, and it also specifies the range of those extreme points.



Note that the maximum marginal manipulation gain $\overline{\Psi'_s}$ depends on $f_s^y(x)$, α_s, and C_s. Given fixed cost C_s, $\overline{\Psi'_s}$ increases as the maximum manipulation benefit $\Delta_s(x_s^*)$ increases and/or α_s decreases (i.e., when there are more unqualified individuals who can manipulate). Given fixed $\Delta_s(x_s^*)$ and α_s, $\overline{\Psi'_s}$ increases as cost decreases (i.e., $P_{C_s}(c)$ is shifted/skewed toward the direction of lower cost). Theorem 22 shows that as compared to non-strategic policy $\widehat{\theta}_s^{\text{UN}}$, strategic policy θ_s^{UN} over(under) selects when a group is majority-(un)qualified.[3] In either case, as shown by Theorem 22, this means $\widehat{\theta}_s^{\text{UN}}$ is always closer to x_s^* (the single peak of $p_s^0(\theta)$) compared to θ_s^{UN}. Therefore, the strategic policy always disincentivizes manipulative behavior, i.e., manipulation probability $p_s^0(\theta_s^{\text{UN}}) < p_s^0(\widehat{\theta}_s^{\text{UN}})$.

Impact on fairness. The characterization of strategic policy $(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ and non-strategic policy $(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$ allows us to further compare them against a given fairness criterion $\mathcal{C}$. Suppose we define

[3] We say $\mathcal{G}_s$ is majority-unqualified (resp. majority-qualified) if $\alpha_s < \delta_u$ (resp. $\alpha_s > \delta_u$). When $u_- = u_+$, $\delta_u = 0.5$, a group is majority-(un)qualified if more than a half of its members are (un)qualified.

the *unfairness* of threshold policy (θ_a, θ_b) as $\mathcal{E}^{\mathcal{C}}(\theta_a, \theta_b) = \mathbb{E}_{X\sim\mathcal{P}_a^{\mathcal{C}}}[\mathbf{1}(x \geq \theta_a)] - \mathbb{E}_{X\sim\mathcal{P}_b^{\mathcal{C}}}[\mathbf{1}(x \geq \theta_b)] = \mathbb{F}_b^{\mathcal{C}}(\theta_b) - \mathbb{F}_a^{\mathcal{C}}(\theta_a)$, where we denote the CDF $\mathbb{F}_s^{\mathcal{C}}(\theta) = \int_{-\infty}^{\theta} \mathcal{P}_s^{\mathcal{C}}(x)dx$. Define the *disadvantaged group* under policy (θ_a, θ_b) as the group with the larger $\mathbb{F}_s^{\mathcal{C}}(\theta_s)$, i.e., the group with the smaller selection rate (DP) or the smaller true positive rate (EqOpt). Define group index $-s := \{a, b\} \setminus s$. Note that we measure unfairness $\mathcal{E}^{\mathcal{C}}(\theta_a, \theta_b)$ over the original feature distributions $f_s^y(x)$ before manipulation. We first identify distributional conditions under which the strategic optimal policy worsens unfairness.

Theorem 23. *If $\alpha_s > \delta_u > \alpha_{-s}$ and $\mathbb{F}_s^{\mathcal{C}}(x_s^*) \leq \mathbb{F}_{-s}^{\mathcal{C}}(x_{-s}^*)$, then strategic policy $(\theta_a^{UN}, \theta_b^{UN})$ has worse unfairness compared to non-strategic $(\widehat{\theta}_a^{UN}, \widehat{\theta}_b^{UN})$, i.e., $\left|\mathcal{E}^{\mathcal{C}}(\theta_a^{UN}, \theta_b^{UN})\right| > \left|\mathcal{E}^{\mathcal{C}}(\widehat{\theta}_a^{UN}, \widehat{\theta}_b^{UN})\right|$, $\mathcal{C} \in \{EqOpt, DP\}$. Moreover, the disadvantaged group under $(\theta_a^{UN}, \theta_b^{UN})$ and $(\widehat{\theta}_a^{UN}, \widehat{\theta}_b^{UN})$ is the same.*

Given the conditions in Theorem 23, $\mathcal{G}_{-s}$ is disadvantaged under non-strategic policy. Because the majority-(un)qualified group $\mathcal{G}_s(\mathcal{G}_{-s})$ is over(under) selected under strategic policy (Theorem 22), $\mathcal{G}_{-s}$ becomes more disadvantaged while $\mathcal{G}_s$ becomes more advantaged, i.e., the unfairness gap is wider under strategic policy. Note that condition $\mathbb{F}_s^{\mathcal{C}}(x_s^*) \leq \mathbb{F}_{-s}^{\mathcal{C}}(x_{-s}^*)$ holds if $f_a^y(x) = f_b^y(x)$. For the DP fairness measure, it holds for any distribution when α_s is sufficiently large or α_{-s} sufficiently small. As shown in Section 6.7, it is also seen in the real world (e.g., FICO data).

We next identify conditions on the manipulation cost, under which strategic policy $(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ can lead to a more equitable outcome or flip the (dis)advantaged group compared to non-strategic $(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$.

Theorem 24. *If $\alpha_a, \alpha_b < \delta_u$ and $\mathbb{F}_{-s}^{\mathcal{C}}(\widehat{\theta}_{-s}^{UN}) > \mathbb{F}_s^{\mathcal{C}}(\widehat{\theta}_s^{UN})$, i.e., $\mathcal{G}_{-s}$ is disadvantaged under non-strategic policy, then given any $\mathcal{G}_{-s}$, there always exists cost C_s for $\mathcal{G}_s$ s.t. $\overline{\Psi_s'}$ is sufficiently large and*

1. *$(\theta_a^{UN}, \theta_b^{UN})$ mitigates the unfairness, i.e., $\left|\mathcal{E}^{\mathcal{C}}(\theta_a^{UN}, \theta_b^{UN})\right| < \left|\mathcal{E}^{\mathcal{C}}(\widehat{\theta}_a^{UN}, \widehat{\theta}_b^{UN})\right|$.*
2. *$(\theta_a^{UN}, \theta_b^{UN})$ flips the disadvantaged group, i.e., $\mathbb{F}_{-s}^{\mathcal{C}}(\theta_{-s}^{UN}) < \mathbb{F}_s^{\mathcal{C}}(\theta_s^{UN})$.*

Because $\alpha_s < \delta_u$, we have $\theta_s^{\text{UN}} > \widehat{\theta}_s^{\text{UN}} > x_s^*$ (by Theorem 22). Moreover, θ_s^{UN} increases as $\Psi_s'(\Delta_s(\theta))$ increases ($P_{C_s}(c)$ is skewed toward the direction of lower cost). Intuitively, as $\mathcal{G}_s$'s manipulation cost decreases, more individuals can afford manipulation; thus a strategic decision maker disincentivizes manipulation by increasing the threshold θ_s^{UN}. For any $\mathcal{G}_{-s}$, as $\mathbb{F}_s^{\mathcal{C}}(\theta_s^{\text{UN}})$ increases, either the unfairness gets mitigated or $\mathbb{F}_s^{\mathcal{C}}(\theta_s^{\text{UN}})$ becomes larger than $\mathbb{F}_{-s}^{\mathcal{C}}(\theta_{-s}^{\text{UN}})$. Proposition 6 in the following considers a special case when $f_a^y(x) = f_b^y(x)$, and gives conditions on $\Psi_s'(\cdot)$ under which $(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ mitigates the unfairness or flips the disadvantaged group when $\mathcal{C} \in \{\text{EqOpt}, \text{DP}\}$.

Proposition 6. *Suppose $f_a^y(x) = f_b^y(x)$, $\alpha_{-s} < \alpha_s < \delta_u$, then $\mathbb{F}^{\mathcal{C}}_{-s}(\widehat{\theta}^{UN}_{-s}) > \mathbb{F}^{\mathcal{C}}_{s}(\widehat{\theta}^{UN}_{s})$, i.e., $\mathcal{G}_{-s}$ is disadvantaged under non-strategic policy. Denote $\Delta(\cdot) = \Delta_a(\cdot) = \Delta_b(\cdot)$. Given any $\mathcal{G}_{-s}$, always there exists manipulation cost C_s for $\mathcal{G}_s$ s.t. $\Psi'_s(\cdot)$ satisfies the followings:*

1. $\frac{\Psi'_s(\Delta(\theta^{UN}_{-s}))-u_+\alpha_s}{\Psi'_{-s}(\Delta(\theta^{UN}_{-s}))-u_+\alpha_{-s}} = \frac{u_-(1-\alpha_s)-u_+\alpha_s}{u_-(1-\alpha_{-s})-u_+\alpha_{-s}}$, *then $\theta^{UN}_a = \theta^{UN}_b$ and $(\theta^{UN}_a, \theta^{UN}_b)$ mitigates unfairness.*

2. $\Psi'_s(\Delta(\eta^{\mathcal{C}}(\theta^{UN}_{-s}))) \geq u_-(1-\alpha_s)$, *then $(\theta^{UN}_a, \theta^{UN}_b)$ flips the disadvantaged group.*

where $(\eta^{\mathcal{C}}(\theta^{UN}_{-s}), \theta^{UN}_{-s})$ satisfies fairness $\mathcal{C}$, i.e., $\eta^{EqOpt}(\theta^{UN}_{-s}) = \theta^{UN}_{-s}$, $\eta^{DP}(\theta^{UN}_{-s}) = (\mathbb{F}^{DP}_s)^{-1}\mathbb{F}^{DP}_{-s}(\theta^{UN}_{-s})$.

Note that above conditions are sufficient. In particular, *case 1* corresponds to the case where the perfect `EqOpt` fairness is attained (i.e., $\mathcal{E}^{\texttt{EqOpt}}(\theta^{\texttt{UN}}_a, \theta^{\texttt{UN}}_b) = 0$) and `DP` fairness is improved (i.e., $\left|\mathcal{E}^{\texttt{DP}}(\theta^{\texttt{UN}}_a, \theta^{\texttt{UN}}_b)\right| < \left|\mathcal{E}^{\texttt{DP}}(\widehat{\theta}^{\texttt{UN}}_a, \widehat{\theta}^{\texttt{UN}}_b)\right|$).

6.6 Impact of Fairness Interventions

In this section, we study how non-strategic and strategic policies are affected by fairness interventions.

Impact of fairness intervention on the non-strategic policy. First, we consider the non-strategic decision maker and compare $(\widehat{\theta}^{\texttt{UN}}_a, \widehat{\theta}^{\texttt{UN}}_b)$ with $(\widehat{\theta}^{\mathcal{C}}_a, \widehat{\theta}^{\mathcal{C}}_b)$, both ignoring strategic manipulation but the latter imposing a fairness criterion. Theorem 25 identifies conditions under which a fairness constrained $(\widehat{\theta}^{\mathcal{C}}_a, \widehat{\theta}^{\mathcal{C}}_b)$ yields *higher* utility from both groups compared to unconstrained $(\widehat{\theta}^{\texttt{UN}}_a, \widehat{\theta}^{\texttt{UN}}_b)$. This is worth noting because had strategic manipulation been absent, policy $(\widehat{\theta}^{\texttt{UN}}_a, \widehat{\theta}^{\texttt{UN}}_b)$ by definition would attain the optimal/highest utility for the decision maker.

Theorem 25. *When $\mathbb{F}^{\mathcal{C}}_s(\widehat{\theta}^{UN}_s) < \mathbb{F}^{\mathcal{C}}_{-s}(\widehat{\theta}^{UN}_{-s})$, i.e., $\mathcal{G}_{-s}$ is disadvantaged under non-strategic optimal policy, then $U_a(\widehat{\theta}^{\mathcal{C}}_a) > U_a(\widehat{\theta}^{UN}_a)$ and $U_b(\widehat{\theta}^{\mathcal{C}}_b) > U_b(\widehat{\theta}^{UN}_b)$ hold under any of the following cases:*

1. *$\alpha_s < \delta_u < \alpha_{-s}$ and $\Psi'_s(\Delta_s(\widehat{\theta}^{\mathcal{C}}_s)) > u_-(1-\alpha_s)$, $\Psi'_{-s}(\Delta_{-s}(\widehat{\theta}^{\mathcal{C}}_{-s})) > u_+\alpha_{-s}$.*

2. *$\alpha_a, \alpha_b > \delta_u$ and $\alpha_s \to \delta_u$ and $\Psi'_a(\Delta_a(\widehat{\theta}^{\mathcal{C}}_a)) > u_+\alpha_a$, $\Psi'_b(\Delta_b(\widehat{\theta}^{\mathcal{C}}_b)) > u_+\alpha_b$.*

3. *$\alpha_a, \alpha_b < \delta_u$ and $\alpha_{-s} \to \delta_u$ and $\Psi'_a(\Delta_a(\widehat{\theta}^{\mathcal{C}}_a)) > u_-(1-\alpha_a)$, $\Psi'_b(\Delta_b(\widehat{\theta}^{\mathcal{C}}_b)) > u_-(1-\alpha_b)$.*

Condition $\alpha_s, \alpha_{-s} \to \delta_u$ means that the qualification rates α_s, α_{-s} are sufficiently close to δ_u. Theorem 25 says that when the marginal manipulation gains of the groups under non-strategic fair policy $(\widehat{\theta}^{\mathcal{C}}_a, \widehat{\theta}^{\mathcal{C}}_b)$ are sufficiently large, $(\widehat{\theta}^{\mathcal{C}}_a, \widehat{\theta}^{\mathcal{C}}_b)$ may outperform $(\widehat{\theta}^{\texttt{UN}}_a, \widehat{\theta}^{\texttt{UN}}_b)$ in terms of both fairness

and utility due to the misalignment of $U_s(\theta)$ and $\widehat{U}_s(\theta)$ caused by manipulation. This means that if the decision maker lacks information or awareness to anticipate manipulative behavior (but which in fact exists), then it would benefit from using a fairness constrained policy $(\widehat{\theta}_a^{\mathcal{C}}, \widehat{\theta}_b^{\mathcal{C}})$ rather than unconstrained policy $(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$.

Impact of fairness intervention on the strategic policy. We now compare $(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ and $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$. We also explore their respective subsequent impact on individuals' manipulative behavior by comparing manipulation probabilities $(p_a^0(\theta_a^{\text{UN}}), p_b^0(\theta_b^{\text{UN}}))$ and $(p_a^0(\theta_a^{\mathcal{C}}), p_b^0(\theta_b^{\mathcal{C}}))$. The goal here is to understand whether fairness intervention can serve as incentives or disincentives for strategic manipulation. According to Theorem 22, $U_s(\theta)$ may have multiple extreme points under strategic manipulation if the group's marginal manipulation gain is sufficiently large. Depending on whether $U_s(\theta)$ has multiple extreme points, different conclusions result as outlined in Theorem 26 below, which identifies conditions under which fairness intervention may increase the manipulation incentive for one group while disincentivizing the other, or it may serve as incentives for both groups.

Theorem 26. *Denote $p_s^{\mathcal{C}} := p_s^0(\theta_s^{\mathcal{C}})$ and $p_s^{UN} := p_s^0(\theta_s^{UN})$. For $\mathcal{C} \in \{\texttt{DP}, \texttt{EqOpt}\}$, we have:*

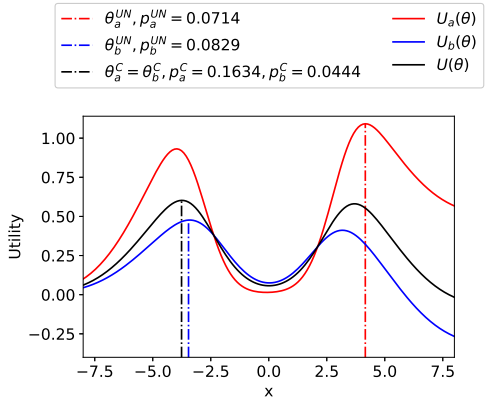
1. *When both $U_a(\theta)$ and $U_b(\theta)$ have unique extreme points, then $\theta_s^{UN} > \theta_s^{\mathcal{C}}$ and $\theta_{-s}^{UN} < \theta_{-s}^{\mathcal{C}}$ must hold. Moreover,*
   - *(i) If $\alpha_s > \delta_u > \alpha_{-s}$, then $\forall \alpha_{-s}$, there exist $\kappa \in (\delta_u, 1)$ and $\tau \in (0,1)$ such that $\forall \alpha_s > \kappa$ and $\forall n_s > \tau$, we have $p_s^{UN} < p_s^{\mathcal{C}}$, $p_{-s}^{UN} > p_{-s}^{\mathcal{C}}$.*
   - *(ii) If $\alpha_a, \alpha_b > \delta_u$ (resp. $\alpha_a, \alpha_b < \delta_u$), then $\forall \alpha_{-s}$, there exists $\kappa \in (\delta_u, 1)$ (resp. $\kappa \in (0, \delta_u)$) such that $\forall\ \alpha_s > \kappa$ (resp. $\alpha_s < \kappa$), we have $p_a^{UN} < p_a^{\mathcal{C}}$, $p_b^{UN} > p_b^{\mathcal{C}}$ or $p_b^{UN} < p_b^{\mathcal{C}}$, $p_a^{UN} > p_a^{\mathcal{C}}$.*
2. *When at least one of $U_a(\theta)$, $U_b(\theta)$ has multiple extreme points, then it is possible that $\forall s \in \{a, b\}$, $\theta_s^{UN} > \theta_s^{\mathcal{C}}$ or $\theta_s^{UN} < \theta_s^{\mathcal{C}}$, i.e., both groups are over/under selected under fair policies. In this case,*
   - *(i) If $\alpha_s > \delta_u > \alpha_{-s}$, we have* $\begin{cases} p_s^{UN} > p_s^{\mathcal{C}}, p_{-s}^{UN} < p_{-s}^{\mathcal{C}} \text{ when } \theta_a^{UN} > \theta_a^{\mathcal{C}}, \theta_b^{UN} > \theta_b^{\mathcal{C}}. \\ p_s^{UN} < p_s^{\mathcal{C}}, p_{-s}^{UN} > p_{-s}^{\mathcal{C}} \text{ when } \theta_a^{UN} < \theta_a^{\mathcal{C}}, \theta_b^{UN} < \theta_b^{\mathcal{C}}. \end{cases}$
   - *(ii) If $\alpha_a, \alpha_b > \delta_u$ (or $\alpha_a, \alpha_b < \delta_u$), we have $p_a^{UN} < p_a^{\mathcal{C}}, p_b^{UN} < p_b^{\mathcal{C}}$ or $p_s^{UN} < p_s^{\mathcal{C}}, p_{-s}^{UN} > p_{-s}^{\mathcal{C}}$.*

When not accounting for strategic manipulation, $\widehat{U}_s(\theta)$ has a unique extreme point, and imposing a fairness constraint results in one group getting under-selected and the other over-selected. In contrast, when the decision maker anticipates strategic manipulation, $U_s(\theta)$ may have multiple

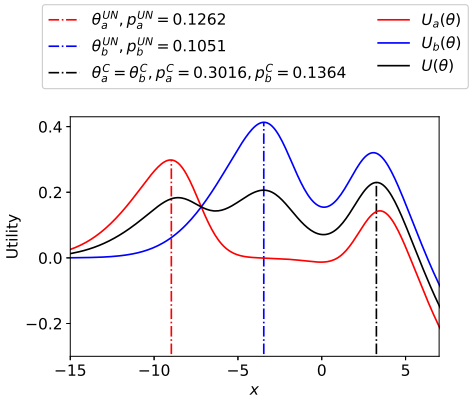
extreme points. One consequence of this difference is that both $\mathcal{G}_a$ and $\mathcal{G}_b$ may be over- or under-selected when fairness is imposed, resulting in more complex incentive relationships. Specifically, if one group is majority-qualified while the other is majority-unqualified, then under-selecting (resp. over-selecting) both groups under fair policies will increase (resp. decrease) the incentives of the former to manipulate, while disincentivizing (resp. incentivizing) the latter (by *2.(i)*); if both groups are majority-(un)qualified, then the fair policy may incentivize both to manipulate (by *2.(ii)*).

If the marginal manipulation gain of both groups are not sufficiently large, i.e., $U_s(\theta)$ has a unique extreme point, then fairness intervention always results in one group getting over-selected and the other under-selected. However, its subsequent impact on incentives may vary depending on $f_s^y(x)$, n_s. Theorem 26 identifies two scenarios under which fair policies incentivize one group (say $\mathcal{G}_s$) while disincentivizing the other ($\mathcal{G}_{-s}$): when $\mathcal{G}_s$ is majority-qualified, $\mathcal{G}_{-s}$ majority-unqualified, and $\mathcal{G}_s$ sufficiently qualified ($\alpha_s > \kappa$) and represented in the entire population ($n_s > \tau$) (by *1.(i)*); or, when both are majority-(un)qualified and $\mathcal{G}_s$ sufficiently (un)qualified (by *1.(ii)*).

An example when both $U_a(\theta)$ and $U_b(\theta)$ have multiple extreme points is shown below where $u_- = u_+$, fairness constraint $\mathcal{C} = \texttt{EqOpt}$.



(a) $f_s^1(x) \sim \mathcal{N}(5,4)$, $f_s^0(x) \sim \mathcal{N}(-5,4), \forall s \in \{a,b\}$, $n_a = 0.3$, $\alpha_a = 0.4, \alpha_b = 0.6$, $C_a \sim \text{Beta}(10,2), C_b \sim \text{Beta}(10,1)$. It shows that $\theta_s^{\mathcal{C}} < \theta_s^{\text{UN}}, \forall s \in \{a,b\}$ and $p_a^{\mathcal{C}} > p_a^{\text{UN}}$, $p_b^{\mathcal{C}} < p_b^{\text{UN}}$.



(b) $f_s^1(x) \sim \mathcal{N}(5,9)$, $f_b^0(x) \sim \mathcal{N}(-5,9)$, $f_a^0(x) \sim \mathcal{N}(-10,9)$, $n_a = 0.5$, $\alpha_a = 0.65, \alpha_b = 0.6$, $C_a \sim \text{Beta}(10,3), C_b \sim \text{Beta}(10,2)$. It shows that $\theta_s^{\mathcal{C}} > \theta_s^{\text{UN}}$ and $p_s^{\mathcal{C}} > p_s^{\text{UN}}, \forall s \in \{a,b\}$.

Because $f_a^1(x) = f_b^1(x)$, under $\texttt{EqOpt}$ fairness, $\theta_a^{\mathcal{C}} = \theta_b^{\mathcal{C}}$ and the total utility $n_a U_a(\theta_a^{\mathcal{C}}) + n_b U_b(\theta_b^{\mathcal{C}})$ can be expressed as a function of $\theta = \theta_a^{\mathcal{C}} = \theta_b^{\mathcal{C}}$. The above two examples show that when $U_a(\theta)$ and $U_b(\theta)$ have multiple extreme points, it's possible that both groups are over (left)/under (right)

selected under strategic fair policies. When $\alpha_b > \delta_u > \alpha_a$ (left), fairness intervention incentivizes $\mathcal{G}_a$ while disincentivizing $\mathcal{G}_b$; when $\alpha_a, \alpha_b > \delta_u$ (right), fairness intervention incentivizes both groups to manipulate. These results are consistent with Theorem 26.

Next, we identify conditions under which fairness intervention can *disincentivize* both groups. Let x_s^{UN} be defined s.t. $\Delta_s(x_s^{\text{UN}}) = \Delta_s(\theta_s^{\text{UN}})$ and $x_s^{\text{UN}} \neq \theta_s^{\text{UN}}$ when $\theta_s^{\text{UN}} \neq x_s^*$. Note that x_s^{UN} is the point at which $p_s^0(x_s^{\text{UN}}) = p_s^0(\theta_s^{\text{UN}})$. Because manipulation probability is single-peaked, fairness intervention incentivizes manipulative behavior of $\mathcal{G}_s$ if $\theta_s^{\mathcal{C}}$ falls between x_s^{UN} and θ_s^{UN}.

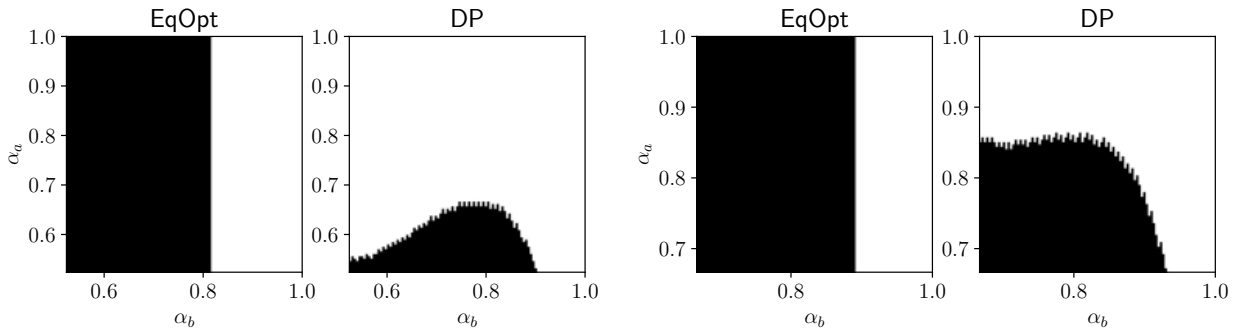
Theorem 27 (Disincentivize both groups)**.** *When both $U_a(\theta)$ and $U_b(\theta)$ have unique extreme points,* $\begin{cases} \text{If } \alpha_a, \alpha_b > \delta_u \text{ and } \mathbb{F}_{-s}^{\mathcal{C}}(x_{-s}^{UN}) < \mathbb{F}_s^{\mathcal{C}}(x_s^*), \text{ then } \exists \kappa > \delta_u \text{ and } \tau \in (0,1) \text{ such that } \forall \alpha_s \in (\delta_u, \kappa) \\ \text{If } \alpha_a, \alpha_b < \delta_u \text{ and } \mathbb{F}_{-s}^{\mathcal{C}}(x_{-s}^{UN}) > \mathbb{F}_s^{\mathcal{C}}(x_s^*), \text{ then } \exists \kappa < \delta_u \text{ and } \tau \in (0,1) \text{ such that } \forall \alpha_s \in (\kappa, \delta_u) \end{cases}$ *and $\forall n_s > \tau$, we have $p_a^{UN} > p_a^{\mathcal{C}}$ and $p_b^{UN} > p_b^{\mathcal{C}}$.*

Note that x_s^* depends on $f_s^y(x)$ and x_{-s}^{UN} is determined by u_-, u_+, $f_{-s}^y(x)$ and α_{-s}. Theorem 27 says that when both groups are majority-(un)qualified, for certain population distributions and $\mathcal{G}_{-s}$, fair policies disincentivize both groups if $\mathcal{G}_s$ is sufficiently unqualified(qualified) and sufficiently represented in the population. For a special Gaussian case, conditions for satisfying $\mathbb{F}_{-s}^{\mathcal{C}}(x_{-s}^{\text{UN}}) \lesseqgtr \mathbb{F}_s^{\mathcal{C}}(x_s^*)$ in Theorem 27 are given in Proposition 7 below.

Proposition 7. *Suppose $f_s^y(x)$ follows Gaussian distribution with mean μ_s^y and variance σ^2. If $0 < \mu_s^1 - \mu_s^0 < \mu_{-s}^1 - \mu_{-s}^0$, i.e., qualified and unqualified individuals from $\mathcal{G}_s$ are less distinguishable than those from $\mathcal{G}_{-s}$, then*

- *$\mathcal{C} =$ `EqOpt`: $\forall \alpha_s > \delta_u$ (resp. $\alpha_s < \delta_u$), there exists $\omega > \delta_u$ (resp. $\omega < \delta_u$) such that $\forall \alpha_{-s} \in [\delta_u, \omega]$ (resp. $\alpha_{-s} \in [\omega, \delta_u]$), conditions $\mathbb{F}_{-s}^{\mathcal{C}}(x_{-s}^{UN}) \lesseqgtr \mathbb{F}_s^{\mathcal{C}}(x_s^*)$ in Theorem 27 hold.*
- *$\mathcal{C} =$ `DP`: if $u_+ < u_-$ (resp. $u_+ > u_-$), then there exist $\omega_1, \omega_2 > \delta_u$ (resp. $\omega_1, \omega_2 < \delta_u$) such that $\forall \alpha_b \in [\delta_u, \omega_1]$ (resp. $\forall \alpha_b \in [\omega_1, \delta_u]$) and $\forall \alpha_a \in [\delta_u, \omega_2]$ (resp. $\forall \alpha_a \in [\omega_2, \delta_u]$), conditions $\mathbb{F}_{-s}^{\mathcal{C}}(x_{-s}^{UN}) \lesseqgtr \mathbb{F}_s^{\mathcal{C}}(x_s^*)$ in Theorem 27 hold.*

Theorems 26 and 27 suggest that the impact of fairness intervention on the individuals' manipulative behavior highly depends on manipulation costs, feature distributions, group qualification and representation. This complexity stems from the misalignment in manipulation probability $p_s^0(\theta)$, utility $U_s(\theta)$, and fairness $\mathcal{C}$. In particular, the manipulation probability of $\mathcal{G}_s$ is single-peaked with maximum at x_s^*, which does not depend on group qualification and representation, but on which



(a) $u_+ : u_- = 1 : 1.1$ (b) $u_+ : u_- = 1 : 2$

Figure 6.3: Examples validating Proposition 7: black region indicates (α_a, α_b) satisfying condition $\mathbb{F}^{\mathcal{C}}_{-s}(x^{\text{UN}}_{-s}) < \mathbb{F}^{\mathcal{C}}_{s}(x^*_s)$ in Theorem 27: $\alpha_a, \alpha_b > \delta_u$, $C_a, C_b \sim \text{Beta}(10, 1)$, $f^y_s(x)$ follows Gaussian distribution with mean μ^y_s and variance σ^2, and $[\mu^0_a, \mu^1_a, \mu^0_b, \mu^1_b] = [-2, 2, -5, 5]$, $\sigma = 4.5$.

the decision maker's total utility depends, as varying α_s and n_s will affect the policies. As a result, depending on which region θ^{UN}_s falls into, i.e., smaller or larger than x^*_s, and how it may change under constraint $\mathcal{C}$, fairness intervention will have different impacts on incentives.

Although Theorems 26 and 27 hold for both `EqOpt` and `DP` fairness, there are scenarios under which they have different impact on incentives. Proposition 8 below further considers a special case when $f^y_a(x) = f^y_b(x)$ and one group is majority-qualified while the other majority-unqualified, in which `EqOpt` never disincentivize both groups while `DP` can disincentivize both.

Proposition 8. *Suppose $f^y_a(x) = f^y_b(x)$, if $\alpha_s > \delta_u > \alpha_{-s}$, then*

- $\forall f^y_s(x)$, $p^{EqOpt}_a < p^{UN}_a$, $p^{EqOpt}_b < p^{UN}_b$ *is unattainable, i.e., `EqOpt` never disincentivize both.*
- $\exists f^y_s(x)$, (α_a, α_b), *and n_a under which $p^{DP}_a < p^{UN}_a$, $p^{DP}_b < p^{UN}_b$, i.e., `DP` may disincentivize both groups.*

6.7 Experiments

We conduct experiments on both a Gaussian synthetic dataset, and the FICO scores dataset [122]. We assume manipulation costs follow either a uniform distribution ($C_s \sim U[0, \overline{c}]$) or a beta distribution ($C_s \sim \text{Beta}(a, b)$, smaller b and larger a lead to larger manipulation costs, Figure 6.4 below illustrates

examples of probability density function $P_{C_s}(z)$ and scaled marginal manipulation gain $\frac{\Psi_s'(z)}{u_-(1-\alpha_s)} = \mathbb{F}_{C_s}(z) + zP_{C_s}(z)$).

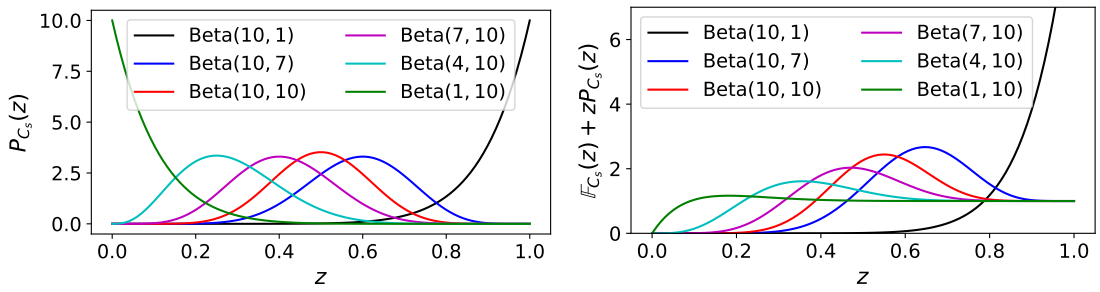


Figure 6.4: Illustration of $P_{C_s}(z)$ and $\mathbb{F}_{C_s}(z) + zP_{C_s}(z)$: $C_s \sim \text{Beta}(a,b)$.

Gaussian data. Suppose $X|Y,S$ is Gaussian distributed. Figure 6.6 shows an example where $f_s^y(x) \sim \mathcal{N}(\mu_s^y, \sigma^2)$ with $[\mu_a^0, \mu_a^1, \mu_b^0, \mu_b^1] = [-2, 2, -5, 5]$, $\sigma = 4.5$, and fairness intervention can serve as disincentive for manipulation for both groups. It shows that $\forall \alpha_b > \delta_u$ satisfying condition $\mathbb{F}_b^{\mathcal{C}}(x_b^{\text{UN}}) < \mathbb{F}_a^{\mathcal{C}}(x_a^*)$, there exist sufficiently small α_a and sufficiently large n_a under which $p_a^0(\theta_a^{\text{UN}}) > p_a^0(\theta_a^{\mathcal{C}})$ and $p_b^0(\theta_b^{\text{UN}}) > p_b^0(\theta_b^{\mathcal{C}})$, i.e., both groups are disincentivized under strategic fair policies. This verifies Theorem 27.

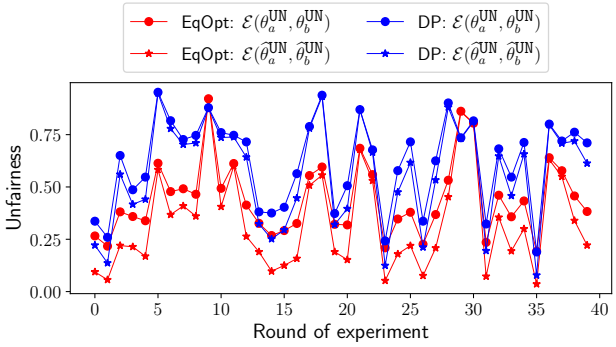


Figure 6.5: Verification of Theorem 23: $C_a \sim \text{Beta}(10,1)$, $C_b \sim \text{Beta}(10,3)$, $u_- = u_+$, $f_s^1(x) \sim \mathcal{N}(5, 5^2)$, $f_s^0(x) \sim \mathcal{N}(-5, 5^2)$, $s = a, b$.

We verify Theorem 23 by conducting 40 rounds of experiment independently. In each round of experiment, (α_a, α_b) is randomly generated with $\alpha_a > \delta_u > \alpha_b$. We consider `EqOpt` (red) or `DP` (blue) as fairness measure. In Figure 6.5, circles and stars represent the unfairness $\mathcal{E}^{\mathcal{C}}(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ and $\mathcal{E}^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$ respectively. It shows that the strategic policy (circles) always worsens the unfairness (both `EqOpt` and `DP`) compared to non-strategic policy (stars), and $\mathcal{G}_b$ is disadvantaged in all scenarios. Varying costs C_s, distributions $f_s^y(x)$, and u_+, u_-, we observe the similar results.

Similarly, we verify Theorem 24 by running 40 rounds of experiments independently. In each round, (α_a, α_b) is randomly generated with $\delta_u > \alpha_a > \alpha_b$. In Figure 6.7, circles that fall below the black dashed line indicate the disadvantaged group being flipped under strategic policy. It shows as $\mathcal{G}_a$'s manipulation cost decreases, unfairness can be mitigated (circles fall below stars) and

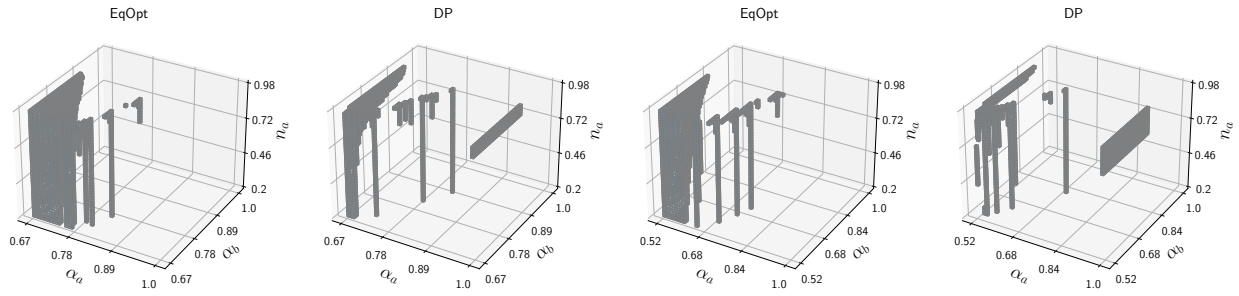
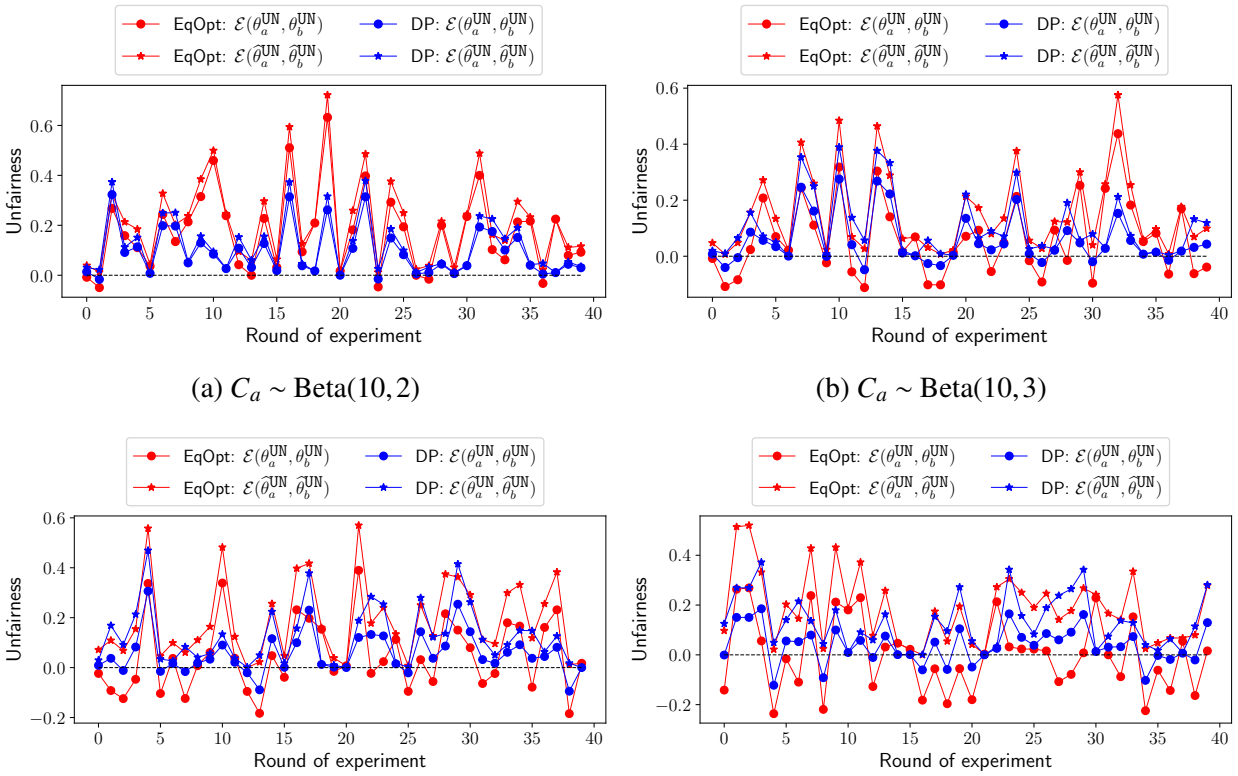


Figure 6.6: $\alpha_a, \alpha_b > \delta_u$, $C_a = C_b \sim \text{Beta}(10,1)$, $\frac{u_+}{u_-} = \frac{1}{2}$ (left), $\frac{u_+}{u_-} = \frac{1}{1.1}$ (right). Grey region indicates $(\alpha_a, \alpha_b, n_a)$ satisfying $\mathbb{F}_b^{\mathcal{C}}(x_b^{\text{UN}}) < \mathbb{F}_a^{\mathcal{C}}(x_a^*)$ in Theorem 27; meanwhile both groups are disincentivized under $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$.



(a) $C_a \sim \text{Beta}(10,2)$

(b) $C_a \sim \text{Beta}(10,3)$

(c) $C_a \sim \text{Beta}(10,4)$

(d) $C_a \sim \text{Beta}(10,5)$

Figure 6.7: $C_b \sim \text{Beta}(10,1)$, $f_s^1(x) \sim \mathcal{N}(5,5^2)$, $f_s^0(x) \sim \mathcal{N}(-5,5^2)$, $s = a, b$.

disadvantaged group can be flipped (circles fall below black dashed line).

Figures 6.8 and 6.9 illustrate the manipulation probabilities of two groups under strategic policy (UN) and strategic fair policy (EqOpt, DP), where $u_+ = u_-$, $C_s \sim \text{Beta}(10,1)$, $f_b^1(x) \sim \mathcal{N}(5,5^2)$, $f_b^0(x) \sim \mathcal{N}(-5,5^2)$, $f_a^1(x) \sim \mathcal{N}(5,4^2)$, $f_a^0(x) \sim \mathcal{N}(-5,4^2)$. Black, blue, red surfaces correspond to $p_s^0(\theta_s^{\text{UN}}) := p_s^{\text{UN}}$, $p_s^0(\theta_s^{\text{DP}}) := p_s^{\text{DP}}$, $p_s^0(\theta_s^{\text{EqOpt}}) := p_s^{\text{EqOpt}}$ respectively. Figure 6.8 shows that when n_a and α_a are sufficiently large, $p_a^{\text{UN}} < p_a^{\mathcal{C}}$ and $p_b^{\text{UN}} > p_b^{\mathcal{C}}$ hold, $\mathcal{C} \in \{\text{EqOpt}, \text{DP}\}$. Figure 6.9 shows when two groups are majority-(un)qualified, $p_a^{\text{UN}} < p_a^{\mathcal{C}}$, $p_b^{\text{UN}} > p_b^{\mathcal{C}}$ or $p_a^{\text{UN}} > p_a^{\mathcal{C}}$, $p_b^{\text{UN}} < p_b^{\mathcal{C}}$ holds as long as one of α_a, α_b is sufficiently large (small).

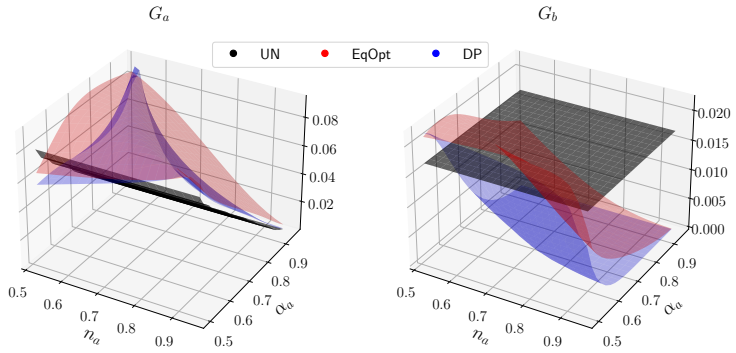
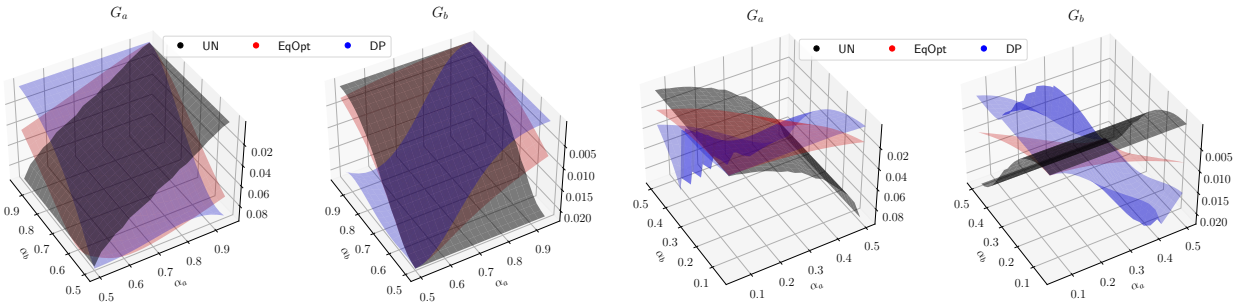


Figure 6.8: Verification of *1(i)* in Theorem 26: $\alpha_b = 0.4$. Varying $\mathcal{G}_a$'s qualification $\alpha_a \in [0.5,1]$ and representation $n_a \in [0.5,1]$, the resulting manipulation probabilities are shown in plots.



(a) $\alpha_a \in [0.5,1]$ and $\alpha_b \in [0.5,1]$

(b) $\alpha_a \in [0,0.5]$ and $\alpha_b \in [0,0.5]$

Figure 6.9: Verification of *1(ii)* in Theorem 26: $n_a = 0.5$. In the left (resp. right), varying two groups' qualification $\alpha_a, \alpha_b > \delta_u$ (resp. $\alpha_a, \alpha_b < \delta_u$) , the resulting manipulation probabilities of two groups are shown in the plots.

FICO scores [122]. FICO scores are widely used in the US to assess an individual's creditworthiness. The is a dataset pre-processed by [57] to generate CDF of scores $\mathbb{F}_{X|S}(x|s)$ and qualification profile $P_{Y|X,S}(1|x,s)$ for different social groups (Caucasian, African-American, Hispanic, Asian). We use these to estimate the conditional feature distribution $f_s^y(x)$ by fitting the simulated data to a Beta distribution. We can see from Figure 6.10 that $\frac{f_s^1(x)}{f_s^0(x)}$ is strictly increasing, it implies that Assumption 11 holds for FICO scores data. This allows us to derive the various equilibrium strategies studied in this chapter. We further calculate repayment rates α_s and proportions n_s. These are summarized in Figures 6.10 & 6.11 and Table 6.1.

Table 6.1: Qualification rate $\alpha_a = P_{Y|S}(1|s)$, conditional feature distributions $f_s^y(x)$, group proportions n_s of four social groups. x_s^* satisfies $f_s^1(x_s^*) = f_s^0(x_s^*)$.

$\mathcal{G}_s$	α_s	$f_s^0(x)$	$f_s^1(x)$	n_s	x_s^*
Caucasian	0.758	Beta(1.23, 12.34)	Beta(2.57, 1.24)	0.7651	0.277
African-American	0.338	Beta(1.18, 15.99)	Beta(1.84, 2.32)	0.1050	0.174
Hispanic	0.570	Beta(1.23, 9.02)	Beta(2.03, 1.90)	0.0845	0.262
Asian	0.804	Beta(0.89, 4.94)	Beta(2.31, 1.38)	0.0454	0.342

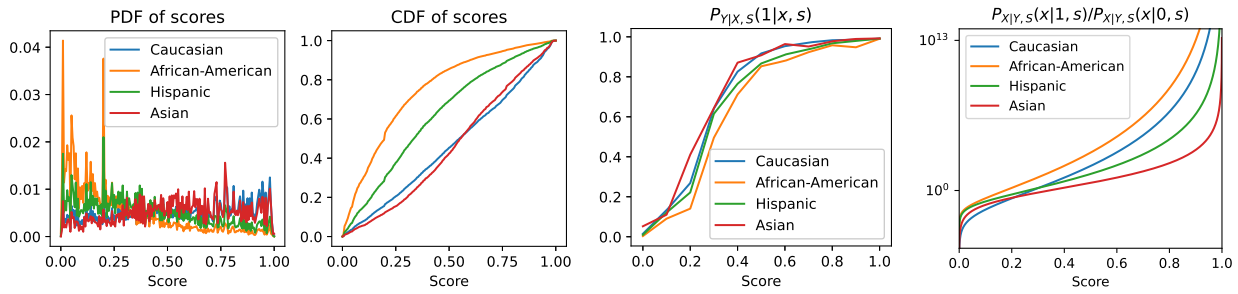


Figure 6.10: Illustration of score PDF/CDF, qualification profiles, and validation of Assumption 11.

We first compare strategic policy $(\theta_a^{\mathrm{UN}}, \theta_b^{\mathrm{UN}})$ and non-strategic policy $(\widehat{\theta}_a^{\mathrm{UN}}, \widehat{\theta}_b^{\mathrm{UN}})$ in terms of their fairness. Let $\mathcal{G}_a$ denote Caucasian, Hispanic or Asian, and $\mathcal{G}_b$ denote African-American. As shown in Table 6.2, $\mathcal{G}_b$ is always disadvantaged compared to other groups, and strategic policy worsens unfairness. When $C_a \neq C_b$, the manipulation cost of $\mathcal{G}_b$ is shifted lower. It further shows that this gets worse when it is less costly for those in $\mathcal{G}_b$ to manipulate their features. Since $\alpha_a > \delta_u > \alpha_b$, this is consistent with Theorem 23.

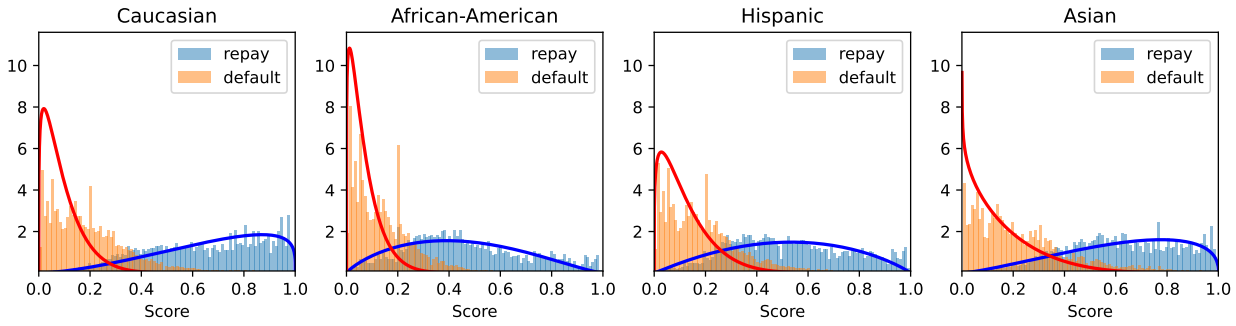


Figure 6.11: Fit Beta distributions to the simulated data to get $f_s^y(x)$.

Table 6.2: Unfairness $\mathcal{E}^{\mathcal{C}}(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ and $\mathcal{E}^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$ for $\mathcal{C} \in \{\texttt{EqOpt}, \texttt{DP}\}$: $\mathcal{G}_b$ = African-American, $u_+ = u_-$, $C_a \sim \text{Beta}(10,2)$ (or $C_a \sim U[0,1]$). When cost $C_a \neq C_b$, $C_b \sim \text{Beta}(10,6)$ (or $C_b \sim U[0,0.5]$).

		EqOpt			DP		
	$\mathcal{G}_a$	strategic		non-strategic	strategic		non-strategic
		$C_a = C_b$	$C_a \neq C_b$		$C_a = C_b$	$C_a \neq C_b$	
Beta	Caucasian	0.355	0.556	0.136	0.611	0.680	0.449
	Hispanic	0.292	0.493	0.034	0.421	0.490	0.242
	Asian	0.333	0.533	0.123	0.634	0.703	0.522
Uniform	Caucasian	0.743	0.871	0.136	0.794	0.838	0.449
	Hispanic	0.722	0.850	0.034	0.684	0.727	0.242
	Asian	0.738	0.866	0.123	0.825	0.868	0.522

Figure 6.12 illustrates how unfairness can be mitigated and how the disadvantaged group can gain advantage under strategic policies. Specifically, let $\mathcal{G}_a, \mathcal{G}_b$ be Hispanic and African-American respectively. We fix $\mathcal{G}_b$ and vary $\mathcal{G}_a$'s manipulation cost. It shows while $\mathcal{G}_b$ is disadvantaged under non-strategic policies ($\mathcal{E}^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}}) > 0$), unfairness can be mitigated under strategic policies as $\mathcal{G}_a$'s manipulation cost decreases, and the disadvantaged group may gain an advantage in the process ($\mathcal{E}^{\mathcal{C}}(\theta_a^{\text{UN}}, \theta_b^{\text{UN}}) < 0$). This is an example of Theorem 24.

According to Theorem 25, under strategic manipulation, non-strategic fair policy $(\widehat{\theta}_a^{\mathcal{C}}, \widehat{\theta}_b^{\mathcal{C}})$ may yield higher utilities from both groups compared to $(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$. We verify this in Table 6.3, in which $\mathcal{G}_a$, $\mathcal{G}_b$ denote Caucasian and Asian groups, respectively, with `EqOpt` as the fairness constraint.

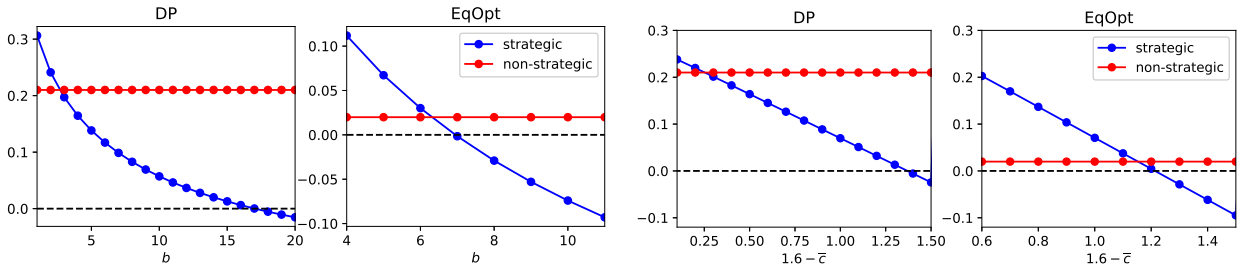


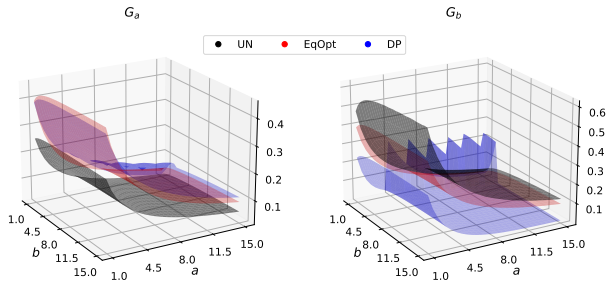
Figure 6.12: Unfairness $\mathcal{E}^{\mathcal{C}}(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ and $\mathcal{E}^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$, $\frac{u_+}{u_-} = \frac{1}{2}$, $\alpha_a, \alpha_b < \delta_u$. Perfect equity is indicated by the black dashed line. $C_b \sim \text{Beta}(10,5)$ and $C_a \sim \text{Beta}(10, b)$ (left), where larger b indicates smaller costs; $C_b \sim U[0,1]$, $C_a \sim U[0, \overline{c}]$ (right).

Table 6.3: $\mathcal{G}_a = \text{Caucasian}(\alpha_a = 0.758)$, $\mathcal{G}_b = \text{Asian}(\alpha_b = 0.804)$, $\mathcal{C} = \texttt{EqOpt}$. The first (resp. second) row corresponds to case 1 (resp. case 2) in Theorem 25.

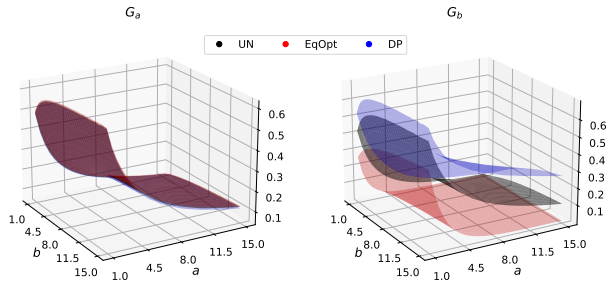
$u_+ : u_-$	δ_u	C_a	C_b	$U_a(\widehat{\theta}_a^{\text{UN}})$	$U_a(\widehat{\theta}_a^{\mathcal{C}})$	$U_b(\widehat{\theta}_b^{\text{UN}})$	$U_b(\widehat{\theta}_b^{\mathcal{C}})$
1 : 4	0.8	Beta(10, 10)	Beta(10, 10)	-0.190	-0.189	0.024	0.034
1 : 3.1	0.756	Beta(10, 1)	Beta(10, 10)	0.396	0.397	0.181	0.201

It illustrates two cases corresponding to cases 1 and 2 in Theorem 25, and $U_a(\widehat{\theta}_a^{\mathcal{C}}) > U_a(\widehat{\theta}_a^{\text{UN}})$, $U_b(\widehat{\theta}_b^{\mathcal{C}}) > U_b(\widehat{\theta}_b^{\text{UN}})$ hold in both cases, i.e., $(\widehat{\theta}_a^{\mathcal{C}}, \widehat{\theta}_b^{\mathcal{C}})$ satisfies fairness and attains higher utility than $(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$.

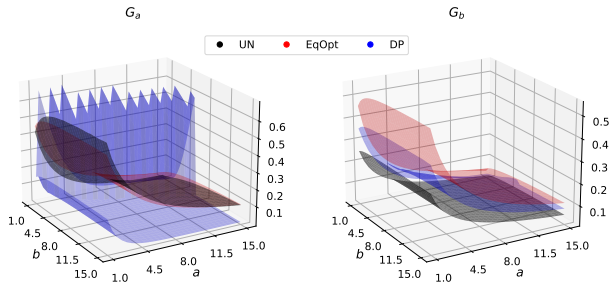
Lastly, we examine how fairness intervention acts as incentives for manipulation. Manipulation probabilities $p_s^0(\theta_s^{\text{UN}})$, $p_s^0(\theta_s^{\texttt{EqOpt}})$, and $p_s^0(\theta_s^{\texttt{DP}})$ are compared under different manipulation costs in Figures 6.13 and 6.14. In Figure 6.13, groups have the same manipulation costs $C_a = C_b \sim \text{Beta}(a, b)$ while in Figure 6.14, $C_a \sim U[0, \overline{c}_a]$ and $C_b \sim U[0, \overline{c}_b]$ are different; $u_- = u_+$ in both cases. Black, red and blue surfaces indicate the manipulation probabilities $p_s^0(\theta_s)$ under $(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$, $(\theta_a^{\texttt{EqOpt}}, \theta_b^{\texttt{EqOpt}})$ and $(\theta_a^{\texttt{DP}}, \theta_b^{\texttt{DP}})$ policies as manipulation costs change. It shows that fairness intervention can incentivize both groups to manipulate (Figure 6.13a), and that `EqOpt` and `DP` may have contrarian impact (Figure 6.13b). Moreover, when there is a significant gap in the two groups' manipulation costs, fairness intervention incentivizes the group with a low manipulation cost while disincentivizing the group with a high manipulation cost (Figure 6.14).



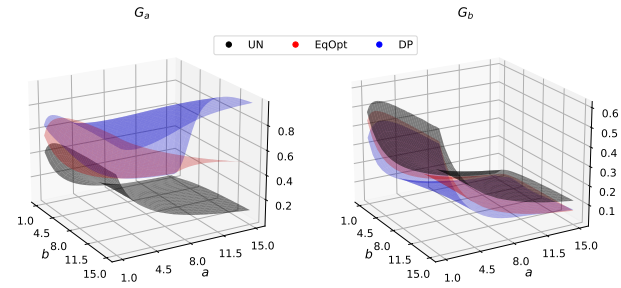
(a) $\mathcal{G}_a$ = African-American; $\mathcal{G}_b$ =Hispanic



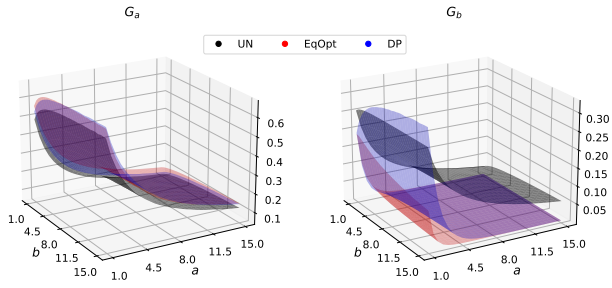
(b) $\mathcal{G}_a$ = Caucasian; $\mathcal{G}_b$ =Asian



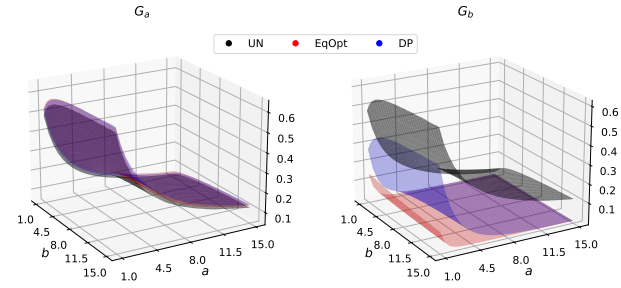
(c) $\mathcal{G}_a$ = Asian; $\mathcal{G}_b$ =African-American



(d) $\mathcal{G}_a$ = Asian; $\mathcal{G}_b$ =Hispanic

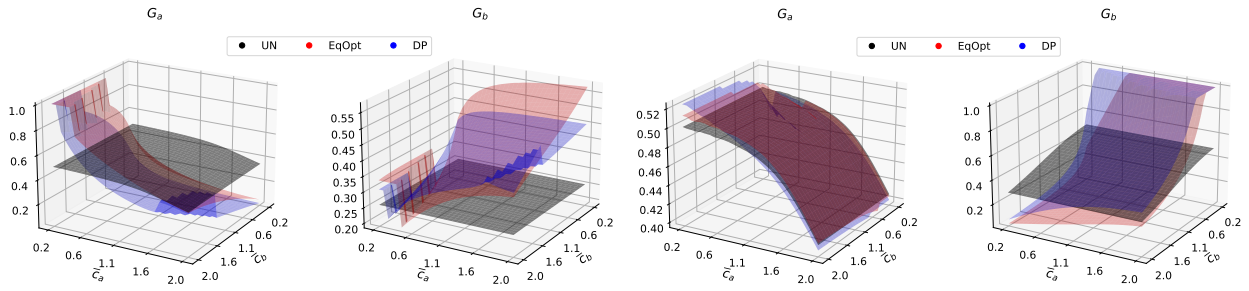


(e) $\mathcal{G}_a$ = Caucasian; $\mathcal{G}_b$ =African-American



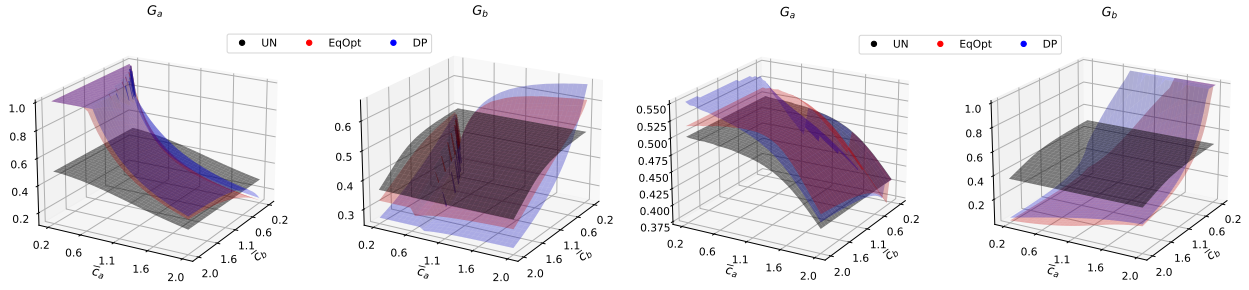
(f) $\mathcal{G}_a$ = Caucasian; $\mathcal{G}_b$ =Hispanic

Figure 6.13: Manipulation probabilities under strategic (fair) policy: $C_a = C_b \sim \text{Beta}(a, b)$, $a \in [1, 15]$, $b \in [1, 15]$.



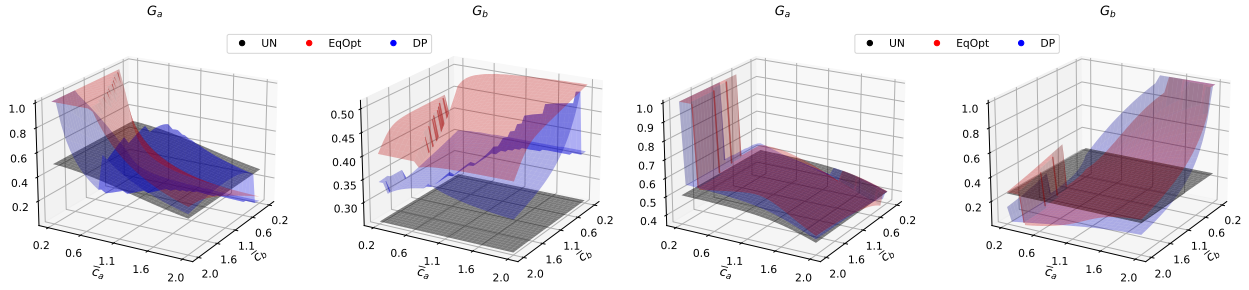
(a) $\mathcal{G}_a$ =Hispanic; $\mathcal{G}_b$ =African-American

(b) $\mathcal{G}_a$ = Caucasian; $\mathcal{G}_b$ =Asian



(c) $\mathcal{G}_a$ =Asian; $\mathcal{G}_b$ =Hispanic

(d) $\mathcal{G}_a$ = Caucasian; $\mathcal{G}_b$ =Hispanic



(e) $\mathcal{G}_a$ =Asian; $\mathcal{G}_b$ =African-American

(f) $\mathcal{G}_a$ = Caucasian; $\mathcal{G}_b$ =African-American

Figure 6.14: Manipulation probabilities under strategic (fair) policy: $C_s \sim U[0, \bar{c}_s], s = a, b$, $\bar{c}_a \in [0.2, 2]$, $\bar{c}_b \in [0.2, 2]$.

CHAPTER 7

Conclusion

7.1 Thesis Summary

As machine learning techniques are increasingly used in domains involving human. It is critical to understand the societal implications of ML systems, and *build* and *use* ML systems responsibly. Toward this end, we studied two important issues, privacy and fairness, in this thesis and investigated the following questions.

(1) When individuals' data are used for building the computational system, how to accomplish the computational goals without violating individual privacy.

(2) When ML systems are used to make decisions about individuals from multiple demographic groups, how do they interact with each other? What are the (long-term) impacts of fairness interventions?

Specifically, we studied (1) in the first part of the thesis. We adopted *differential privacy* as notion of privacy and developed a number of randomized algorithms for two types of computations: distributed learning and sequential computations. Because the same/correlated data is repeatedly used during these computations, balancing the trade-off between outcome accuracy and individual privacy is challenging. It was shown that our algorithms can achieve a higher accuracy than the existing algorithms under the same privacy guarantee. These algorithms are developed based on two ideas to improve this privacy-accuracy tradeoff: (a) reuse intermediate computations to reduce the total information leakage so that less noise is needed to guarantee a certain level of privacy, and (b) improve algorithmic robustness so that more noise can be accommodated to enhance privacy without jeopardizing too much accuracy.

In the second part of the thesis, we investigated problem (2) by studying fairness problems with human in feedback loops. We constructed three types of dynamics to capture the interplay between individuals and ML systems: (a) *participation dynamics*: ML decisions affect individuals' retention in the system which further affects the group representations in future dataset. (b) *qualification dynamics:* ML decisions affect individuals' qualifications in the system which further affects the rate of positive class in future dataset. (c) *strategic dynamics:* individuals by knowing how decisions are made given input can manipulate their features strategically which affects ML system given decision maker being aware of such manipulative behavior. Under each dynamics, we conducted equilibrium analysis to understand the impacts ML system and individuals each have on the other, and explored the role of fairness interventions by studying how these equilibria are affected when a certain fairness constraint is imposed in ML systems.

Specifically, when studying participation and qualification dynamics, we show that fairness intervention that intends to protect the disadvantaged group may actually cause harm to the disadvantaged group by exacerbating the group disparity, i.e., the disparity in group qualifications (under qualification dynamics) and disparity in group representations (under participation dynamics), in the long run. We identified conditions under which such worsening of fairness may happen and proposed potential mitigating solutions: for participation dynamics, we formulated an optimization problem for finding a proper fairness notion that can sustain group representations over time; for qualification dynamics, we suggested policy/transition interventions that can either lead to a more equitable equilibrium or improve qualifications of both groups in the long run.

For strategic dynamics where individuals strategically manipulate features, we examined the impacts of decision maker's awareness of strategic manipulation and impacts of fairness interventions. We identified conditions under which being aware of strategic manipulation can improve or worsen the fairness, and conditions under which fairness interventions can serve as incentives or disincentives for strategic manipulation. When decision maker lacks awareness to anticipate manipulative behavior, we identified conditions under which decision maker also benefits from the fairness interventions.

The societal implications of these quantitative results are as follows. Firstly, our results may help policymakers (e.g., companies, banks, governments, etc.) in their decision making process by highlighting the potential pitfalls of commonly used static fairness criteria and providing guidance on how to design effective interventions (e.g., based on Chapter 5, giving community support to social groups to increase their transitions and long-term qualifications) that can avoid such

unintended consequences and result in positive long-term societal impacts. Secondly, our results may be useful to research in fields outside of the computer science community. For example, the experiments in Chapter 5 have shown consistent findings with literature in social sciences [52, 119]. Although these empirical results are obtained using simulated dynamics due to a lack of real datasets, they may provide insights and theoretical supports for research in other fields. Lastly, while this work is limited to binary decisions, the main take-away can be applied in other applications such as computer vision, natural language processing, etc., using more complicated classifiers such as DNN. We hope that our work will encourage researchers in these domains to similarly consider discrimination risks when developing techniques, and raise awareness that static fairness constraint may not suffice and long-term fairness cannot be designed in a vacuum without considering the human element. We thus emphasize the importance of performing real-time measurements and developing proper fair classifiers from dynamic datasets.

7.2 Limitations & Future Directions

We also want to point out the limitations and the potential extensions.

For the studies on private distributed learning, we proposed R-ADMM to improve tradeoff between individual privacy and accuracy. Note that R-ADMM violates monotonicity property of ADMM because of the linearized approximation introduced in even iterations. It is worthwhile to study the impact of linearization and analyse the convergence rate of R-ADMM. Although experiments on real-world data showed that the fluctuation (non-monotonicity) of R-ADMM induced by linearization decreases over time, and the convergence rates between R-ADMM and original ADMM were compared empirically, a theoretical explanation is needed to better understand R-ADMM. As such, one potential future direction is to conduct robust analysis and theoretically prove the fluctuation converges to 0.

For the studies on fairness with human in feedback loops, some limitations and extensions are as follows. Firstly, we partitioned the entire population based on a single binary demographic attribute and considered two demographic groups in our work. This cannot adequately capture the heterogeneity of the population, and there can be multiple non-binary demographic attributes. Secondly, the dynamics we formulated are simplified models. The individual behaviors in real-world are much more complicated and can vary across individuals from the same demographic group. For instance, in qualification dynamics, we use a set of transitions to capture individuals' abilities

to improve/maintain future qualifications, and our analysis and conclusions rely on this set of values; in strategic dynamics, we assumed individuals are fully rational who always take actions that maximize its own utility and there is a constraint on manipulation cost. The experiments are also conducted on uniform/beta distributed cost. However, in practice, quantities of transitions and manipulation costs can be extremely hard to measure due to the complexity of human behaviors and environmental factors. Thirdly, we considered the case where individuals' labels/qualification states are binary while in many applications qualification is continuous on a spectrum. As a result, transitions in qualification dynamics can not be captured by four quantifies and all individuals (not only the unqualified ones) may have an incentive to manipulate under strategic dynamics. Fourthly, our results indeed are sensitive to the models and depends on certain assumptions. For example, as discussed in Section 5.7, the change in transition types or feature distributions could lead to significant change in consequences. Moreover, Chapter 5 separately studies two scenarios that causes *natural inequality*, i.e., either transitions or feature distributions are demographic-variant. However, in practice it's likely that both are demographic-variant. This can complicates the dynamical model significantly and adds more uncertainty. It is not clear how our results are robust to the perturbation in dynamics. Lastly, due to the lack of dynamic datasets, our experiments are performed over static real-world datasets with simulated dynamics.

Some extensions to tackle the aforementioned limitations are as follows. One potential direction is to extend models from binary qualifications/demographic groups to non-binary cases. For strategic dynamics, it is worthwhile to study a more generalized scenario where all individuals have incentives to manipulate, and consider partially rational individuals who instead of taking actions that maximize their utilities take sub-optimal actions. Another extension is to conduct sensitivity analysis to understand the robustness of our results.

In addition to the extensions of current studies, there are several long-term future directions.

The Intersection Between Privacy and Fairness in ML. In this thesis, we studied privacy and fairness issues separately. Indeed, there is a strong connection between them. It is interesting to study the impact of one on the other (e.g., whether achieving privacy helps improve fairness and vice versa) [88]. On the other hand, sometimes achieving one societal constraint may add difficulties to satisfy another. For instance, it becomes more difficult to develop fair ML models when protected attributes (e.g., race, gender) are private and unobservable. Building upon the relations between privacy and fairness, we will also develop ML systems that simultaneously satisfy both.

Learning Human Behavioral Models. The second part of this thesis has highlighted the importance of understanding human behaviors in building ML systems. As mentioned, three types of dynamics studied in the thesis are simplified models, it is critical to learn interpretable human behavioral models using ML techniques via empirical studies. One potential future direction is to develop online crowdsourcing platforms or survey sites to collect dynamic data from people and then use ML algorithms to train a human behavioral model. Such a model is essential for building ML systems with long-term social benefits. It may help advance ML research towards a more interpretable domain and open up the possibility of understanding the causal relationships of human-generated data.

Ethical Issues From Multiple Disciplinary Perspectives. Ethical issue such as fairness or privacy itself is complicated and controversial. It is critical to consider them from multiple disciplinary perspectives such as economics, social sciences, law, etc. For example, there is no universal notion of fairness and the proper notions are context dependent. Finding the proper notions that best capture human perception, especially in dynamic environment, and notions that aligned with law and policies is critical for building ML system with social benefits.

APPENDIX A

Private ADMM-Based Distributed Algorithms

A.1 Proof of Simplifying ADMM [45]

By KKT condition of (2.5), there is:

$$0 = \lambda_{ij}^{b}(t) - \lambda_{ij}^{a}(t) + \eta(2w_{ij}(t+1) - f_i(t+1) - f_j(t+1))$$

Implies:

$$w_{ij}(t+1) = \frac{1}{2\eta}(\lambda_{ij}^{a}(t) - \lambda_{ij}^{b}(t)) + \frac{1}{2}(f_i(t+1) + f_j(t+1)) \tag{A.1}$$

Plug (A.1) into (2.6)(2.7):

$$\lambda_{ij}^{a}(t+1) = \frac{1}{2}(\lambda_{ij}^{a}(t) + \lambda_{ij}^{b}(t)) + \frac{\eta}{2}(f_i(t+1) - f_j(t+1)) \tag{A.2}$$

$$\lambda_{ij}^{b}(t+1) = \frac{1}{2}(\lambda_{ij}^{b}(t) + \lambda_{ij}^{a}(t)) + \frac{\eta}{2}(f_i(t+1) - f_j(t+1)) \tag{A.3}$$

If initialize $\lambda_{ij}^{a}(0) = \lambda_{ij}^{b}(0)$ to be zero vectors for all node pairs (i, j), (A.2)(A.3) imply that $\lambda_{ij}^{a}(t) = \lambda_{ij}^{b}(t)$ and $\lambda_{ji}^{k}(t) = -\lambda_{ij}^{k}(t), k \in \{a, b\}$ will hold for all t. (A.1) becomes:

$$w_{ij}(t+1) = \frac{1}{2}(f_i(t+1) + f_j(t+1)) \tag{A.4}$$

Let $\lambda_{ij}(t) = \lambda_{ij}^{a}(t) = \lambda_{ij}^{b}(t)$, (2.6)(2.7) can be simplified as:

$$\lambda_{ij}(t+1) = \lambda_{ij}(t) + \frac{\eta}{2}(f_i(t+1) - f_j(t+1)) \tag{A.5}$$

Plug (A.4) into the augmented Lagrangian (2.3) to simplify it:

$$\begin{aligned} L_\eta(\{f_i\},\{w_{ij},\lambda_{ij}^k\}) &= \sum_{i=1}^N O(f_i,D_i) + \sum_{i=1}^N \sum_{j\in\mathscr{V}_i} (\lambda_{ij}(t))^T (f_i - f_j) \\ &+ \sum_{i=1}^N \sum_{j\in\mathscr{V}_i} \frac{\eta}{2} (\|f_i - \frac{1}{2}(f_i(t)+f_j(t))\|_2^2) + \sum_{i=1}^N \sum_{j\in\mathscr{V}_i} \frac{\eta}{2} (\|\frac{1}{2}(f_i(t)+f_j(t)) - f_j\|_2^2) \end{aligned} \tag{A.6}$$

Since $\sum_{i=1}^N \sum_{j\in\mathscr{V}_i} \lambda_{ij}(t) f_j = \sum_{i=1}^N \sum_{j\in\mathscr{V}_i} \lambda_{ji}(t) f_i$ and $\lambda_{ij}(t) = -\lambda_{ji}(t)$, the second term in (A.6) can be simplified:

$$\sum_{i=1}^N \sum_{j\in\mathscr{V}_i} (\lambda_{ij}(t))^T (f_i - f_j) = 2 \sum_{i=1}^N \sum_{j\in\mathscr{V}_i} (\lambda_{ij}(t))^T f_i$$

The last term can be expressed as:

$$\sum_{i=1}^N \sum_{j\in\mathscr{V}_i} \frac{\eta}{2} (\|\frac{1}{2}(f_i(t)+f_j(t)) - f_j\|_2^2) = \sum_{i=1}^N \sum_{j\in\mathscr{V}_i} \frac{\eta}{2} (\|\frac{1}{2}(f_i(t)+f_j(t)) - f_i\|_2^2)$$

Therefore, (A.6) is simplified as:

$$L_\eta(\{f_i\},\{w_{ij},\lambda_{ij}^k\}) = \sum_{i=1}^N O(f_i,D_i) + 2\sum_{i=1}^N \sum_{j\in\mathscr{V}_i} \lambda_{ij}(t)^T f_i + \sum_{i=1}^N \sum_{j\in\mathscr{V}_i} \eta(\|f_i - \frac{1}{2}(f_i(t)+f_j(t))\|_2^2) \tag{A.7}$$

Define $\lambda_i(t) = \sum_{j\in\mathscr{V}_i} \lambda_{ij}(t)$. Based on (A.5)(A.7), the original ADMM updates (2.4)-(2.7) are simplified as:

$$f_i(t+1) = \underset{f_i}{\operatorname{argmin}}\ O(f_i,D_i) + 2\lambda_i(t)^T f_i + \eta \sum_{j\in\mathscr{V}_i} \|f_i - \frac{1}{2}(f_i(t)+f_j(t))\|_2^2$$

$$\lambda_i(t+1) = \lambda_i(t) + \frac{\eta}{2} \sum_{j\in\mathscr{V}_i} (f_i(t+1) - f_j(t+1))$$

A.2 Proof of Theorem 1

Subtract (2.20) from (2.27) and (2.21) from (2.28):

$$\begin{aligned}&\nabla\hat{O}(\hat{f}(t+1),D_{all})-\nabla\hat{O}(\hat{f}^*,D_{all})+\sqrt{D-A}(Y(t+1)-Y^*)\\&+(W(t+1)-\theta I)(D-A)\hat{f}(t+1)+W(t+1)(D+A)(\hat{f}(t+1)-\hat{f}(t))=\mathbf{0}_{N\times d}\end{aligned}\tag{A.8}$$

$$Y(t+1)=Y(t)+\theta\sqrt{D-A}(\hat{f}(t+1)-\hat{f}^*)\tag{A.9}$$

By convexity of $O(f_i,D_i)$, for any f_i^1 and f_i^2, there is:

$$(f_i^1-f_i^2)^T(\nabla O(f_i^1,D_i)-\nabla O(f_i^2,D_i))\geq 0$$

Let $\langle\cdot,\cdot\rangle_F$ be frobenius inner product of two matrices, there is:

$$\langle\hat{f}(t+1)-\hat{f}^*,\nabla\hat{O}(\hat{f}(t+1),D_{all})-\nabla\hat{O}(\hat{f}^*,D_{all})\rangle_F\geq 0$$

Substitute $\nabla\hat{O}(\hat{f}(t+1),D_{all})-\nabla\hat{O}(\hat{f}^*,D_{all})$ from (A.8):

$$\begin{aligned}0\ \leq\ &\langle\hat{f}(t+1)-\hat{f}^*,-\sqrt{D-A}(Y(t+1)-Y^*)\rangle_F\\&+\langle\hat{f}(t+1)-\hat{f}^*,-(W(t+1)-\theta I)(D-A)\hat{f}(t+1)\rangle_F\\&+\langle\hat{f}(t+1)-\hat{f}^*,-W(t+1)(D+A)(\hat{f}(t+1)-\hat{f}(t))\rangle_F\end{aligned}\tag{A.10}$$

Consider the right hand side of (A.10). Since $D-A$ is symmetric and PSD, $\sqrt{D-A}$ is also a symmetric matrix and by (A.9),

$$\begin{aligned}&\langle\hat{f}(t+1)-\hat{f}^*,-\sqrt{D-A}(Y(t+1)-Y^*)\rangle_F\\=\ &\langle-\sqrt{D-A}(\hat{f}(t+1)-\hat{f}^*),(Y(t+1)-Y^*)\rangle_F\\=\ &-\langle\frac{1}{\theta}(Y(t+1)-Y(t)),Y(t+1)-Y^*\rangle_F\end{aligned}\tag{A.11}$$

Rearrange (A.10) and use $(D-A)\hat{f}^*=\mathbf{0}_{N\times d}$

$$\begin{aligned}0\ \geq\ &\langle Z(t+1)-Z^*,J(t+1)(Z(t+1)-Z(t))\rangle_F\\&+\langle\hat{f}(t+1)-\hat{f}^*,(W(t+1)-\theta I)(D-A)(\hat{f}(t+1)-\hat{f}^*)\rangle_F\end{aligned}\tag{A.12}$$

Suppose $\eta_i(t) \geq \theta$ for all t, i, i.e., the diagonal matrix $W(t) - \theta I \geq 0$ for all t. Since $D - A \geq 0$, whose eigenvalues are all non-negative, the eigenvalues of $(W(t+1) - \theta I)(D - A)$ are thus also non-negative, i.e., $(W(t+1) - \theta I)(D - A) \geq 0$. Then for the second term of the RHS of (A.12), there is:

$$\langle \hat{f}(t+1) - \hat{f}^*, (W(t+1) - \theta I)(D - A)(\hat{f}(t+1) - \hat{f}^*) \rangle_F \geq 0$$

Therefore,

$$\langle Z(t+1) - Z^*, J(t+1)(Z(t+1) - Z(t)) \rangle_F \leq 0 \tag{A.13}$$

To simplify the notation, for a matrix X, let $\|X\|_J^2 = \langle X, JX \rangle_F$, then (A.13) can be represented as:

$$\frac{1}{2}\|Z(t+1) - Z^*\|_{J(t+1)}^2 + \frac{1}{2}\|Z(t+1) - Z(t)\|_{J(t+1)}^2 - \frac{1}{2}\|Z(t) - Z^*\|_{J(t+1)}^2 \leq 0$$

implies

$$\begin{aligned} \|Z(t+1) - Z(t)\|_{J(t+1)}^2 \quad \leq \quad & -\|Z(t+1) - Z^*\|_{J(t+1)}^2 + \|Z(t) - Z^*\|_{J(t)}^2 \\ & + \|Z(t) - Z^*\|_{J(t+1)}^2 - \|Z(t) - Z^*\|_{J(t)}^2 \end{aligned} \tag{A.14}$$

Suppose $\eta_i(t+1) \geq \eta_i(t)$ for all t and i, i.e., the diagonal matrix $W(t+1) - W(t) \geq 0$ for all t. Since $D + A \geq 0$, implies $(W(t+1) - W(t))(D + A) \geq 0$. Let $U = \sup_{i,t,k} |(f_i(t) - f_c^*)_k| \in \mathbb{R}$ be the finite upper bound of all nodes i, all iterations t and all components k, then

$$\begin{aligned} & \|Z(t) - Z^*\|_{J(t+1)}^2 - \|Z(t) - Z^*\|_{J(t)}^2 \\ = \quad & \mathrm{Tr}((Z(t) - Z^*)^T (J(t+1) - J(t))(Z(t) - Z^*)) \\ = \quad & \mathrm{Tr}((\hat{f}(t) - \hat{f}^*)^T (W(t+1) - W(t))(D + A)(\hat{f}(t) - \hat{f}^*)) \\ & \leq U^2(\|\mathbf{ones}(N,d)\|_{W(t+1)(D+A)}^2 - \mathbf{ones}(N,d)\|_{W(t)(D+A)}^2) \end{aligned} \tag{A.15}$$

where $\mathbf{ones}(N,d)$ is all one's matrix of size $N \times d$. By (A.14)(A.15):

$$\begin{aligned} \|Z(t+1) - Z(t)\|_{J(t+1)}^2 \leq \|Z(t) - Z^*\|_{J(t)}^2 - \|Z(t+1) - Z^*\|_{J(t+1)}^2 \\ + U^2(\|\mathbf{ones}(N,d)\|_{W(t+1)(D+A)}^2 - \|\mathbf{ones}(N,d)\|_{W(t)(D+A)}^2) \end{aligned} \tag{A.16}$$

Sum up (A.16) over t from 0 to $+\infty$ leads to:

$$\begin{aligned}\sum_{t=0}^{+\infty} \|Z(t+1)-Z(t)\|^2_{J(t+1)} \leq \|Z(0)-Z^*\|^2_{J(0)} - \|Z(+\infty)-Z^*\|^2_{J(+\infty)} \\ +U^2(\|\mathbf{ones}(N,d)\|^2_{W(+\infty)(D+A)} - \|\mathbf{ones}(N,d)\|^2_{W(0)(D+A)})\end{aligned} \tag{A.17}$$

Since $\eta_i(t) < +\infty$, the RHS of (A.17) is finite, implies that $\lim_{t\to+\infty} \|Z(t+1)-Z(t)\|^2_{J(t+1)} = 0$ must hold.

By the definition of $Z(t)$, $J(t)$ and $\|X\|^2_J = \langle X, JX\rangle_F$, the following must hold

$$\lim_{t\to+\infty} \|\hat{f}(t+1)-\hat{f}(t)\|^2_{W(t+1)(D+A)} = 0 \tag{A.18}$$

$$\lim_{t\to+\infty} \|Y(t+1)-Y(t)\|^2_F = 0 \tag{A.19}$$

(A.19) shows that $Y(t)$ converges to a stationary point Y^s, along with (2.28) imply $\lim_{t\to+\infty} \sqrt{D-A}\hat{f}(t+1) = 0$. Since $\text{Null}(\sqrt{D-A}) = c\mathbf{1}$, $\hat{f}(t+1)$ must lie in the subspace spanned by $\mathbf{1}$ as $t \to \infty$. To satisfy (A.18), either of the following two statements must hold:

- $\lim_{t\to+\infty}(\hat{f}(t+1)-\hat{f}(t)) = \mathbf{0}_{N\times d}$
- $\lim_{t\to+\infty} W(t+1)(D+A)\mathbf{1} = \lim_{t\to+\infty} W(t+1)A\mathbf{1} + \lim_{t\to+\infty}\sum_{i=1}^{N}\eta_i(t+1)V_i = \mathbf{0}_{N\times 1}$

Since $\eta_i(t) \geq \theta > 0$ for all t, implies $\lim_{t\to+\infty}\sum_{i=1}^{N}\eta_i(t+1)V_i > 0$. The second statement can never be true because all elements of A and $W(t+1)$ are non-negative. Hence, $\hat{f}(t)$ should also converge to a stationary point $\hat{f}^s$.

Now show that the stationary point $(Y^s, \hat{f}^s)$ is $(Y^*, \hat{f}^*)$.

Take limit of both sides of (2.27) (2.28), substitute $\hat{f}^s, Y^s$ yields

$$\nabla\hat{O}(\hat{f}^s, D_{all}) + \sqrt{D-A}Y^s + (W(t+1)-\theta I)(D-A)\hat{f}^s = \mathbf{0}_{N\times d} \tag{A.20}$$

$$\sqrt{D-A}\hat{f}^s = \mathbf{0}_{N\times d} \tag{A.21}$$

By (A.21), (A.20) turns into:

$$\nabla\hat{O}(\hat{f}^s, D_{all}) + \sqrt{D-A}Y^s = \mathbf{0}_{N\times d} \tag{A.22}$$

Compare (A.21)(A.22) with (2.20)(2.21) in Lemma 1 and observe that $(Y^s, \hat{f}^s)$ satisfies the optimality condition (2.20)(2.21) and is thus the optimal point. Therefore, $f(t)$ converges to $\hat{f}^*$ and $Y(t)$ converges to Y^*.

A.3 Proof of Theorem 2

According to the Assumption 3 that $O(f_i, D_i)$ is strongly convex and has Lipschitz continues gradients for all $i \in \mathcal{N}$, define diagonal matrices $D_m = \mathbf{diag}([m_1; m_2; \cdots; m_N]) \in \mathbb{R}^{N\times N}$ and $D_M = \mathbf{diag}([M_1^2; M_2^2; \cdots; M_N^2]) \in \mathbb{R}^{N\times N}$, (2.30) yield:

$$\langle \hat{f}^1 - \hat{f}^2, \nabla\hat{O}(\hat{f}^1, D_{all}) - \nabla\hat{O}(\hat{f}^2, D_{all})\rangle_F \geq \langle \hat{f}^1 - \hat{f}^2, D_m(\hat{f}^1 - \hat{f}^2)\rangle_F \tag{A.23}$$

$$\|\nabla\hat{O}(\hat{f}^1, D_{all}) - \nabla\hat{O}(\hat{f}^2, D_{all})\|_F^2 \leq \langle \hat{f}^1 - \hat{f}^2, D_M(\hat{f}^1 - \hat{f}^2)\rangle_F \tag{A.24}$$

Since for any $\mu > 1$ and any matrices C_1, C_2 with the same dimensions, there is:

$$\|C_1 + C_2\|_F^2 \leq \mu\|C_1\|_F^2 + \frac{\mu}{\mu - 1}\|C_2\|_F^2$$

From (A.8), there is:

$$\begin{aligned}
\|\sqrt{D-A}(Y(t+1) - Y^*)\|_F^2 \quad \leq \quad & \mu\|\nabla\hat{O}(\hat{f}(t+1), D_{all}) - \nabla\hat{O}(\hat{f}^*, D_{all}) \\
& + W(t+1)(D+A)(\hat{f}(t+1) - \hat{f}(t))\|_F^2 \\
& + \frac{\mu}{\mu-1}\|(W(t+1) - \theta I)(D-A)\hat{f}(t+1)\|_F^2 \\
\leq \quad & \frac{\mu^2}{\mu-1}\|\nabla\hat{O}(\hat{f}(t+1), D_{all}) - \nabla\hat{O}(\hat{f}^*, D_{all})\|_F^2 \\
& + \mu^2\|W(t+1)(D+A)(\hat{f}(t+1) - \hat{f}(t))\|_F^2 \\
& + \frac{\mu}{\mu-1}\|(W(t+1) - \theta I)(D-A)\hat{f}(t+1)\|_F^2
\end{aligned} \tag{A.25}$$

Let $\sigma_{\min}(\cdot)$, $\sigma_{\max}(\cdot)$ denote the smallest nonzero singular value and the largest singular value of a matrix respectively.

For any matrices C_1, C_2, let $C_1 = U\Sigma V^T$ be SVD of C_1, there is:

$$\|C_1C_2\|_F^2 \leq \sigma_{\max}(C_1)\|C_2\|_{C_1^T}^2$$

$$\sigma_{\min}(C_1)^2\|C_2\|_F^2 \leq \|C_1C_2\|_F^2 \leq \sigma_{\max}(C_1)^2\|C_2\|_F^2$$

Denote

$$\bar{\sigma}_{\max}(t+1) = \sigma_{\max}((W(t+1)-\theta I)(D-A))$$

$$\bar{\sigma}_{\min}(t+1) = \sigma_{\min}((W(t+1)-\theta I)(D-A))$$

$$\tilde{\sigma}_{\max}(t+1) = \sigma_{\max}(W(t+1)(D+A))$$

Using (A.24) and $(D-A)\hat{f}^* = 0$, (A.25) is turned into:

$$\frac{1}{\theta}\|Y(t+1)-Y^*\|_F^2 \leq \frac{\mu^2}{\theta\sigma_{\min}(D-A)(\mu-1)}\|\hat{f}(t+1)-\hat{f}^*\|_{D_M}^2$$
$$+\frac{\mu^2\tilde{\sigma}_{\max}(t+1)}{\theta\sigma_{\min}(D-A)}\|\hat{f}(t+1)-\hat{f}(t)\|_{W(t+1)(D+A)}^2 + \frac{\mu\bar{\sigma}_{\max}(t+1)^2}{\theta\sigma_{\min}(D-A)(\mu-1)}\|(\hat{f}(t+1)-\hat{f}^*)\|_F^2$$

Adding $\|\hat{f}(t+1)-\hat{f}^*\|_{W(t+1)(D+A)}^2$ at both sides leads to:

$$\begin{aligned}\|Z(t+1)-Z^*\|_{J(t+1)}^2 \leq{}& \frac{\mu^2\tilde{\sigma}_{\max}(t+1)}{\theta\sigma_{\min}(D-A)}\|\hat{f}(t+1)-\hat{f}(t)\|_{W(t+1)(D+A)}^2 \\ &+\|\hat{f}(t+1)-\hat{f}^*\|_{\frac{\mu^2 D_M+\mu\bar{\sigma}_{\max}(t+1)^2\mathbf{I}_N}{\theta\sigma_{\min}(D-A)(\mu-1)}+W(t+1)(D+A)}^2\end{aligned} \tag{A.26}$$

Since

$$\frac{\delta(t+1)\mu^2\tilde{\sigma}_{\max}(t+1)}{\theta\sigma_{\min}(D-A)} \leq 1 \tag{A.27}$$

and

$$\delta(t+1)\left(\frac{\mu\bar{\sigma}_{\max}(t+1)^2\mathbf{I}_N+\mu^2 D_M}{\theta\sigma_{\min}(D-A)(\mu-1)}+W(t+1)(D+A)\right) \leq 2(W(t+1)-\theta I)(D-A)+2D_m \tag{A.28}$$

It implies from (A.26) that:

$$\begin{aligned} & \delta(t+1)\|Z(t+1)-Z^*\|^2_{J(t+1)} \\ \leq\ & \|\hat{f}(t+1)-\hat{f}(t)\|^2_{W(t+1)(D+A)} + \|\hat{f}(t+1)-\hat{f}^*\|^2_{2(W(t+1)-\theta I)(D-A)+2D_m} \\ \leq\ & \|Z(t+1)-Z(t)\|^2_{J(t+1)} + \|\hat{f}(t+1)-\hat{f}^*\|^2_{2(W(t+1)-\theta I)(D-A)+2D_m} \end{aligned} \tag{A.29}$$

Substituting $\hat{f}^1$ with $\hat{f}(t+1)$ and $\hat{f}^2$ with $\hat{f}^*$ and the gradient difference from (A.8) in (A.23) leads to:

$$\begin{aligned} &\langle \hat{f}(t+1)-\hat{f}^*, \sqrt{D-A}(Y(t+1)-Y^*)\rangle_F + \langle \hat{f}(t+1)-\hat{f}^*, W(t+1)(D+A)(\hat{f}(t+1)-\hat{f}(t))\rangle_F \\ &+\langle \hat{f}(t+1)-\hat{f}^*, (W(t+1)-\theta I)(D-A)\hat{f}(t+1)\rangle_F \leq -\langle \hat{f}(t+1)-\hat{f}^*, D_m(\hat{f}(t+1)-\hat{f}^*)\rangle_F \end{aligned}$$

Similar to the proof of Theorem 1, using the definition of $Z(t+1)$, Z^*, $J(t+1)$ and $(D-A)\hat{f}^* = 0$, there is:

$$\begin{aligned} &\|Z(t+1)-Z^*\|^2_{J(t+1)} \leq -\|Z(t+1)-Z(t)\|^2_{J(t+1)} \\ &+\|Z(t)-Z^*\|^2_{J(t+1)} - \|\hat{f}(t+1)-\hat{f}^*\|^2_{2D_m+2(W(t+1)-\theta I)(D-A)} \end{aligned} \tag{A.30}$$

Sum up (A.29) and (A.30) gives:

$$(1+\delta(t+1))\|Z(t+1)-Z^*\|^2_{J(t+1)} \leq \|Z(t)-Z^*\|^2_{J(t+1)}$$

Let $m_o = \min_{i\in\mathcal{N}}\{m_i\}$, $M_O = \max_{i\in\mathcal{N}}\{M_i\}$. One $\delta(t+1)$ that satisfies (B.5) and (A.28) could be:

$$\min\{\frac{\theta\sigma_{\min}(D-A)}{\mu^2\tilde{\sigma}_{\max}(t+1)}, \frac{2m_o+2\bar{\sigma}_{\min}(t+1)}{\frac{\mu^2M_O^2+\mu\bar{\sigma}_{\max}(t+1)^2}{\theta\sigma_{\min}(D-A)(\mu-1)}+\tilde{\sigma}_{\max}(t+1)}\}$$

A.4 Proof of Theorem 3

By convexity of $O(f_i, D_i)$, $(f_i^1 - f_i^2)^T(\nabla O(f_i^1, D_i) - \nabla O(f_i^2, D_i)) \geq 0$ holds $\forall\ f_i^1, f_i^2$. Let $\langle\cdot,\cdot\rangle_F$ be frobenius inner product of two matrices, there is:

$$\langle \hat{f}(t+1)-\hat{f}^*, \nabla\hat{O}(\hat{f}(t+1), D_{all}) - \nabla\hat{O}(\hat{f}^*, D_{all})\rangle_F \geq 0$$

According to (2.42)(2.22) and (2.43), substitute $\nabla\hat{O}(\hat{f}(t+1), D_{all}) - \nabla\hat{O}(\hat{f}^*, D_{all})$ and add an extra term $W(t+1)(D+A)\tilde{D}(t)^{-1}(\nabla\hat{O}(\hat{f}^*, D_{all}) + 2\Lambda^*) = \mathbf{0}_{N\times d}$, implies Eqn. (A.31).

$$\begin{aligned}
& \langle \hat{f}(t+1) - \hat{f}^*, -W(t+1)(D+A)\tilde{D}(t)^{-1}(\nabla\hat{O}(\hat{f}(t), D_{all}) - \nabla\hat{O}(\hat{f}^*, D_{all})) \\
+ \quad & (I + W(t+1)(D+A)\tilde{D}(t)^{-1})(2\Lambda^* - 2\Lambda(t+1)) \\
+ \quad & W(t+1)(D+A)\tilde{D}(t)^{-1}(2\Lambda(t+1) - 2\Lambda(t)) - W(t+1)(D+A)(\hat{f}(t+1) - \hat{f}(t)) \\
- \quad & W(t+1)(D+A)\tilde{D}^{-1}W(t)(D-A)\hat{f}(t)\rangle_F \geq 0 \,. \qquad \text{(A.31)}
\end{aligned}$$

To simplify the notation, for a matrix X, let $\|X\|_J^2 = \langle X, JX\rangle_F$ and $(X)^+$ be the pseudo inverse of X. Define:

$$\begin{aligned}
G_1(t+1) &= W(t+1)(D+A)\tilde{D}(t)^{-1}W(t)(D-A) \,; \\
G_2(t+1) &= (W(t+1)(D-A))^+ \cdot (I + W(t+1)(D+A)\tilde{D}(t)^{-1}) \,.
\end{aligned}$$

Use (2.43)(2.23) and the fact that $\langle A, JB\rangle_F = \langle J^T A, B\rangle_F$, we have

$$\begin{aligned}
& \langle \hat{f}(t+1) - \hat{f}^*, W(t+1)(D+A)\tilde{D}(t)^{-1}(2\Lambda(t+1) - 2\Lambda(t)) \qquad \text{(A.32)} \\
& -W(t+1)(D+A)\tilde{D}(t)^{-1}W(t)(D-A)\hat{f}(t)\rangle_F \\
= \quad & \langle \hat{f}(t+1) - \hat{f}^*, W(t+1)(D+A)\tilde{D}(t)^{-1}W(t)(D-A)(\hat{f}(t+1) - \hat{f}(t))\rangle_F \\
& + \langle \hat{f}(t+1) - \hat{f}^*, W(t+1)(D+A)\tilde{D}(t)^{-1}(W(t+1) - W(t))(D-A)(\hat{f}(t+1) - f^*)\rangle_F \\
= \quad & \frac{1}{2}\|\hat{f}(t+1) - \hat{f}^*\|^2_{G_1(t+1)} + \frac{1}{2}\|\hat{f}(t+1) - \hat{f}(t)\|^2_{G_1(t+1)} - \frac{1}{2}\|\hat{f}(t) - \hat{f}^*\|^2_{G_1(t+1)} \\
& + \langle \hat{f}(t+1) - \hat{f}^*, W(t+1)(D+A)\tilde{D}(t)^{-1}(W(t+1) - W(t))(D-A)(\hat{f}(t+1) - f^*)\rangle_F \,;
\end{aligned}$$

and

$$\begin{aligned}
& \langle \hat{f}(t+1) - \hat{f}^*, (I + W(t+1)(D+A)\tilde{D}(t)^{-1})(2\Lambda^* - 2\Lambda(t+1))\rangle_F \\
= \quad & \langle (W(t+1)(D-A))^+(2\Lambda(t+1) - 2\Lambda(t)), \\
& (I + W(t+1)(D+A)\tilde{D}(t)^{-1})(2\Lambda^* - 2\Lambda(t+1))\rangle_F \\
= \quad & \frac{1}{2}\|2\Lambda^* - 2\Lambda(t)\|^2_{G_2(t+1)} - \frac{1}{2}\|2\Lambda^* - 2\Lambda(t+1)\|^2_{G_2(t+1)} - \frac{1}{2}\|2\Lambda(t+1) - 2\Lambda(t)\|^2_{G_2(t+1)} \,;
\end{aligned}$$

and

$$\begin{aligned} & \langle \hat{f}(t+1)-\hat{f}^*, -W(t+1)(D+A)(\hat{f}(t+1)-\hat{f}(t))\rangle_F \\ = \; & \frac{1}{2}\|\hat{f}(t)-\hat{f}^*\|^2_{W(t+1)(D+A)} - \frac{1}{2}\|\hat{f}(t+1)-\hat{f}^*\|^2_{W(t+1)(D+A)} \\ & -\frac{1}{2}\|\hat{f}(t)-\hat{f}(t+1)\|^2_{W(t+1)(D+A)} \, . \end{aligned} \tag{A.33}$$

Let $\sqrt{X}$ denote the square root of a symmetric positive semi-definite (PSD) matrix X that is also symmetric PSD. Eqn. (A.34) holds,

$$\begin{aligned} & \langle \hat{f}(t+1)-\hat{f}^*, -W(t+1)(D+A)\tilde{D}(t)^{-1}(\nabla\hat{O}(\hat{f}(t),D_{all})-\nabla\hat{O}(\hat{f}^*,D_{all}))\rangle_F \qquad (A.34) \\ = \; & \langle \hat{f}(t+1)-\hat{f}(t)+\hat{f}(t)-\hat{f}^*, -W(t+1)(D+A)\tilde{D}(t)^{-1}(\nabla\hat{O}(\hat{f}(t),D_{all})-\nabla\hat{O}(\hat{f}^*,D_{all}))\rangle_F \\ \leq \; & \langle \hat{f}(t)-\hat{f}(t+1), W(t+1)(D+A)\tilde{D}(t)^{-1}(\nabla\hat{O}(\hat{f}(t),D_{all})-\nabla\hat{O}(\hat{f}^*,D_{all}))\rangle_F \\ = \; & \langle W(t+1)(D+A)\sqrt{\tilde{D}(t)^{-1}}(\hat{f}(t)-\hat{f}(t+1)), \sqrt{\tilde{D}(t)^{-1}}(\nabla\hat{O}(\hat{f}(t),D_{all})-\nabla\hat{O}(\hat{f}^*,D_{all}))\rangle_F \, . \end{aligned}$$

where the inequality uses the facts that $O(f_i, D_i)$ is convex for all i and that the matrix $W(t+1)(D+A)\tilde{D}(t)^{-1}$ is positive definite.

According to (2.34) in Assumption 4, define the matrix $D_M = \mathbf{diag}([M_1^2; M_2^2; \cdots; M_N^2]) \in \mathbb{R}^{N\times N}$, it implies

$$\|\nabla\hat{O}(\hat{f}^1, D_{all}) - \nabla\hat{O}(\hat{f}^2, D_{all})\|_F^2 \leq \langle \hat{f}^1 - \hat{f}^2, D_M(\hat{f}^1 - \hat{f}^2)\rangle_F$$

Since $\langle A, B\rangle_F \leq \frac{1}{L}\|A\|_F^2 + \frac{L}{4}\|B\|_F^2$ holds for any $L > 0$, there is:

$$\begin{aligned} (A.34) \quad \leq \; & \frac{1}{L}\|W(t+1)(D+A)\sqrt{\tilde{D}(t)^{-1}}(\hat{f}(t)-\hat{f}(t+1))\|_F^2 \\ & +\frac{L}{4}\|\sqrt{\tilde{D}(t)^{-1}}(\nabla\hat{O}(\hat{f}(t),D_{all})-\nabla\hat{O}(\hat{f}^*,D_{all}))\|_F^2 \qquad (A.35) \\ \leq \; & \frac{1}{L}\|(\hat{f}(t)-\hat{f}(t+1))\|^2_{W(t+1)(D+A)\tilde{D}(t)^{-1}W(t+1)(D+A)} + \frac{L}{4\sigma_{\min}(\tilde{D}(t))}\|\hat{f}^*-\hat{f}(t)\|^2_{D_M} \end{aligned}$$

where $\sigma_{\max}(\cdot)$, $\sigma_{\min}(\cdot)$ denote the largest and smallest singular value of a matrix respectively. Since for any $\mu > 1$ and any matrices C_1, C_2, J with the same dimensions, there is $\|C_1 + C_2\|_J^2 \leq$

$\mu\|C_1\|_J^2 + \frac{\mu}{\mu-1}\|C_2\|_J^2$. which implies:

$$\begin{aligned}\|\hat{f}^* - \hat{f}(t)\|_{D_M}^2 &= \|\hat{f}^* - \hat{f}(t+1) + \hat{f}(t+1) - \hat{f}(t)\|_{D_M}^2 \\ &\leq \mu\|\hat{f}^* - \hat{f}(t+1)\|_{D_M}^2 + \frac{\mu}{\mu-1}\|\hat{f}(t+1) - \hat{f}(t)\|_{D_M}^2\end{aligned}$$

Plug into (A.35) and use (2.43)(2.23) gives Eqn. (A.36).

$$\begin{aligned}(\text{A.34}) \quad &\leq \frac{1}{L}\|(\hat{f}(t) - \hat{f}(t+1))\|^2_{W(t+1)(D+A)\tilde{D}(t)^{-1}W(t+1)(D+A)} \\ &\quad + \frac{L}{4\sigma_{\min}(\tilde{D}(t))}(\mu\|\hat{f}^* - \hat{f}(t+1)\|_{D_M}^2 + \frac{\mu}{\mu-1}\|\hat{f}(t+1) - \hat{f}(t)\|_{D_M}^2) \\ &= \frac{1}{2}\|(\hat{f}(t) - \hat{f}(t+1))\|^2_{\frac{2}{L}W(t+1)(D+A)\tilde{D}(t)^{-1}W(t+1)(D+A) + \frac{L\mu}{2\sigma_{\min}(\tilde{D}(t))(\mu-1)}D_M} \\ &\quad + \frac{1}{2}\|2\Lambda(t+1) - 2\Lambda(t)\|^2_{\frac{L\mu}{2\sigma_{\min}(\tilde{D}(t))}((W(t+1)(D-A))^+)^2 D_M} \qquad (\text{A.36})\end{aligned}$$

Combine (A.32)(A.33)(A.33)(A.36), (A.31) becomes Eqn. (A.37).

$$\begin{aligned}&\frac{1}{2}\|\hat{f}(t) - \hat{f}(t+1)\|^2_{W(t+1)(D+A)-G_1(t+1)} \qquad (\text{A.37}) \\ &-\frac{1}{2}\|(\hat{f}(t) - \hat{f}(t+1))\|^2_{\frac{2}{L}W(t+1)(D+A)\tilde{D}(t)^{-1}W(t+1)(D+A) + \frac{L\mu}{2\sigma_{\min}(\tilde{D}(t))(\mu-1)}D_M} \\ &+\frac{1}{2}\|2\Lambda(t+1) - 2\Lambda(t)\|^2_{G_2(t+1)} - \frac{1}{2}\|2\Lambda(t+1) - 2\Lambda(t)\|^2_{\frac{L\mu}{2\sigma_{\min}(\tilde{D}(t))}((W(t+1)(D-A))^+)^2 D_M} \\ \leq\ &\frac{1}{2}\|\hat{f}(t+1) - \hat{f}^*\|^2_{G_1(t+1)} - \frac{1}{2}\|\hat{f}(t) - \hat{f}^*\|^2_{G_1(t+1)} + \frac{1}{2}\|2\Lambda^* - 2\Lambda(t)\|^2_{G_2(t+1)} \\ &-\frac{1}{2}\|2\Lambda^* - 2\Lambda(t+1)\|^2_{G_2(t+1)} + \frac{1}{2}\|\hat{f}(t) - \hat{f}^*\|^2_{W(t+1)(D+A)} - \frac{1}{2}\|\hat{f}(t+1) - \hat{f}^*\|^2_{W(t+1)(D+A)} \\ &+\langle \hat{f}(t+1) - \hat{f}^*, W(t+1)(D+A)\tilde{D}(t)^{-1}(W(t+1) - W(t))(D-A)(\hat{f}(t+1) - f^*)\rangle_F\end{aligned}$$

Suppose the following two conditions hold for all t under some constants $L > 0$ and $\mu > 1$:

$$\begin{aligned}
(i) \quad I + W(t+1)(D+A)\tilde{D}(t)^{-1} &\succ \frac{L\mu}{2\sigma_{\min}(\tilde{D}(t))}(W(t+1)(D-A))^{+}D_M \ ; \\
(ii) \quad W(t+1)(D+A) &\succ W(t+1)(D+A)\tilde{D}(t)^{-1}\Big(W(t)(D-A) \\
&\quad + \frac{2}{L}W(t+1)(D+A)\Big) + \frac{L\mu}{2\sigma_{\min}(\tilde{D}(t))(\mu-1)}D_M \ .
\end{aligned}$$

Substitute $G_1(t+1)$ and $G_2(t+1)$, define $R_1(t+1)$ and $R_2(t+1)$ as (A.38)(A.39). By conditions *(i)(ii)*, both $R_1(t+1)$ and $R_2(t+1)$ are positive definite.

$$\begin{aligned}
R_1(t+1) &= W(t+1)(D+A) - G_1(t+1) - \frac{2}{L}W(t+1)(D+A)\tilde{D}(t)^{-1}W(t+1)(D+A) \\
&\quad - \frac{L\mu}{2\sigma_{\min}(\tilde{D}(t))(\mu-1)}D_M \succ \mathbf{0}_{N\times N} \ ; \qquad \text{(A.38)} \\
R_2(t+1) &= -\frac{L\mu}{2\sigma_{\min}(\tilde{D}(t))}((W(t+1)(D-A))^{+})^2 D_M + G_2(t+1) \succ \mathbf{0}_{N\times N} \ . \qquad \text{(A.39)}
\end{aligned}$$

Eqn. (A.37) becomes Eqn. (A.40).

$$\begin{aligned}
&\frac{1}{2}\|\hat{f}(t) - \hat{f}(t+1)\|^2_{R_1(t+1)} + \frac{1}{2}\|2\Lambda(t+1) - 2\Lambda(t)\|^2_{R_2(t+1)} \qquad \text{(A.40)} \\
\leq\ &\frac{1}{2}\|\hat{f}(t+1) - \hat{f}^*\|^2_{G_1(t+1)} - \frac{1}{2}\|\hat{f}(t) - \hat{f}^*\|^2_{G_1(t+1)} + \frac{1}{2}\|2\Lambda^* - 2\Lambda(t)\|^2_{G_2(t+1)} \\
&-\frac{1}{2}\|2\Lambda^* - 2\Lambda(t+1)\|^2_{G_2(t+1)} + \frac{1}{2}\|\hat{f}(t) - \hat{f}^*\|^2_{W(t+1)(D+A)} - \frac{1}{2}\|\hat{f}(t+1) - \hat{f}^*\|^2_{W(t+1)(D+A)} \\
&+\langle \hat{f}(t+1) - \hat{f}^*, W(t+1)(D+A)\tilde{D}(t)^{-1}(W(t+1) - W(t))(D-A)(\hat{f}(t+1) - f^*)\rangle_F
\end{aligned}$$

Since $W(t+1)$, $W(t)$ and $\tilde{D}(t)$ are all diagonal matrices of the same size, define new diagonal matrix $D_1^{new}(t+1)$ with $D_1^{new}(t+1)_{ii} = \frac{\eta_i(t+1)\eta_i(t)}{2\eta_i(t)V_i+\gamma}$, then $G_1(t+1)$ can be rewritten as:

$$G_1(t+1) = D_1^{new}(t+1)(D+A)(D-A).$$

Consider

$$\begin{aligned}\frac{1}{2}\|\hat{f}(t+1)-\hat{f}^*\|^2_{G_1(t+1)}-\frac{1}{2}\|\hat{f}(t)-\hat{f}^*\|^2_{G_1(t+1)} &= \frac{1}{2}\|\hat{f}(t+1)-\hat{f}^*\|^2_{G_1(t+1)}-\frac{1}{2}\|\hat{f}(t)-\hat{f}^*\|^2_{G_1(t)}\\ &\quad+\frac{1}{2}\|\hat{f}(t)-\hat{f}^*\|^2_{G_1(t)}-\frac{1}{2}\|\hat{f}(t)-\hat{f}^*\|^2_{G_1(t+1)}\end{aligned}$$

If $\eta_i(t+1)\geq\eta_i(t)$, $\forall t,i$, then $D_1^{new}(t+1)_{ii}\geq D_1^{new}(t)_{ii}$. Therefore, $G_1(t+1)-G_1(t)\geq 0$. Let $U_1=\sup_{i,t,k}|(f_i(t)-f_c^*)_k|\in\mathbb{R}$ be the finite upper bound over all components k, all nodes i and all iterations t, then

$$\begin{aligned}\frac{1}{2}\|\hat{f}(t)-\hat{f}^*\|^2_{G_1(t)}-\frac{1}{2}\|\hat{f}(t)-\hat{f}^*\|^2_{G_1(t+1)} &= \frac{1}{2}\mathrm{Tr}((\hat{f}(t)-\hat{f}^*)^T(G_1(t)-G_1(t+1))(\hat{f}(t)-\hat{f}^*))\\ &\leq \frac{1}{2}U_1^2(\|\mathbf{1}_{N\times d}\|^2_{G_1(t+1)}-\|\mathbf{1}_{N\times d}\|^2_{G_1(t)})\end{aligned}$$

where $\mathbf{1}_{N\times d}$ is the matrix of size N by d with 1 on all the entries. Therefore,

$$\begin{aligned}\frac{1}{2}\|\hat{f}(t+1)-\hat{f}^*\|^2_{G_1(t+1)}-\frac{1}{2}\|\hat{f}(t)-\hat{f}^*\|^2_{G_1(t+1)} &\leq \frac{1}{2}\|\hat{f}(t+1)-\hat{f}^*\|^2_{G_1(t+1)}-\frac{1}{2}\|\hat{f}(t)-\hat{f}^*\|^2_{G_1(t)}\\ &\quad+\frac{1}{2}U_1^2(\|\mathbf{1}_{N\times d}\|^2_{G_1(t+1)}-\|\mathbf{1}_{N\times d}\|^2_{G_1(t)})\end{aligned}$$

Similarly, $(W(t+1)-W(t))(D+A)\geq 0$ holds if $\eta_i(t+1)\geq\eta_i(t)$, $\forall t,i$, and the following holds.

$$\begin{aligned}&\frac{1}{2}\|\hat{f}(t)-\hat{f}^*\|^2_{W(t+1)(D+A)}-\frac{1}{2}\|\hat{f}(t+1)-\hat{f}^*\|^2_{W(t+1)(D+A)}\\ &\leq\frac{1}{2}\|\hat{f}(t)-\hat{f}^*\|^2_{W(t)(D+A)}-\frac{1}{2}\|\hat{f}(t+1)-\hat{f}^*\|^2_{W(t+1)(D+A)}\\ &\qquad+\frac{1}{2}U_1^2(\|\mathbf{1}_{N\times d}\|^2_{W(t+1)(D+A)}-\|\mathbf{1}_{N\times d}\|^2_{W(t)(D+A)})\end{aligned}$$

Similarly, if $\eta_i(t+1)\geq\eta_i(t)$, $\forall t,i$, $G_2(t)-G_2(t+1)\geq 0$. Let $U_2=\sup_{i,t,k}|(\lambda_i(t)-\lambda_i^*)_k|\in\mathbb{R}$ be the

finite upper bound over all components k, all nodes i and all iterations t, there is:

$$\begin{aligned} & \frac{1}{2}\|2\Lambda^* - 2\Lambda(t)\|^2_{G_2(t+1)} - \frac{1}{2}\|2\Lambda^* - 2\Lambda(t+1)\|^2_{G_2(t+1)} \\ \leq\ & \frac{1}{2}\|2\Lambda^* - 2\Lambda(t)\|^2_{G_2(t)} - \frac{1}{2}\|2\Lambda^* - 2\Lambda(t+1)\|^2_{G_2(t+1)} \\ +\ & \frac{1}{2}U_2^2(\|\mathbf{1}_{N\times d}\|^2_{G_2(t)} - \|\mathbf{1}_{N\times d}\|^2_{G_2(t+1)}) \end{aligned}$$

If $\eta_i(t+1) \geq \eta_i(t)$, $\forall t, i$, let $\overline{\sigma}_{\max} = \max_t \sigma_{\max}(W(t+1)(D+A)\tilde{D}(t)^{-1}(D-A))$, then there is:

$$\begin{aligned} & \langle \hat{f}(t+1) - \hat{f}^*, W(t+1)(D+A)\tilde{D}(t)^{-1} \cdot (W(t+1) - W(t))(D-A)(\hat{f}(t+1) - f^*)\rangle_F \\ \leq\ & \overline{\sigma}_{\max} U_1^2(\|\mathbf{1}_{N\times d}\|^2_{W(t+1)} - \|\mathbf{1}_{N\times d}\|^2_{W(t)}) \end{aligned}$$

Sum up (A.40) over t from 0 to $+\infty$ leads to:

$$\begin{aligned} & \sum_{t=0}^{\infty}\{\|\hat{f}(t) - \hat{f}(t+1)\|^2_{R_1(t+1)} + \|2\Lambda(t+1) - 2\Lambda(t)\|^2_{R_2(t+1)}\} \\ \leq\ & \|\hat{f}(0) - \hat{f}^*\|^2_{W(0)(D+A)} - \|\hat{f}(+\infty) - \hat{f}^*\|^2_{W(+\infty)(D+A)} + \|\hat{f}(+\infty) - \hat{f}^*\|^2_{G_1(+\infty)} \\ & -\|\hat{f}(0) - \hat{f}^*\|^2_{G_1(0)} + \|2\Lambda^* - 2\Lambda(0)\|^2_{G_2(0)} - \|2\Lambda^* - 2\Lambda(+\infty)\|^2_{G_2(+\infty)} \\ & + U_1^2(\|\mathbf{1}_{N\times d}\|^2_{G_1(+\infty)} - \|\mathbf{1}_{N\times d}\|^2_{G_1(0)}) + U_1^2(\|\mathbf{1}_{N\times d}\|^2_{W(+\infty)(D+A)} - \|\mathbf{1}_{N\times d}\|^2_{W(0)(D+A)}) \\ & + U_2^2(\|\mathbf{1}_{N\times d}\|^2_{G_2(0)} - \|\mathbf{1}_{N\times d}\|^2_{G_2(+\infty)}) + 2\overline{\sigma}_{\max} U_1^2(\|\mathbf{1}_{N\times d}\|^2_{W(+\infty)} - \|\mathbf{1}_{N\times d}\|^2_{W(0)}) \end{aligned} \tag{A.41}$$

The RHS of (A.41) is finite, implies that $\lim_{t\to\infty}\{\|\hat{f}(t) - \hat{f}(t+1)\|^2_{R_1(t+1)} + \|2\Lambda(t+1) - 2\Lambda(t)\|^2_{R_2(t+1)}\} = 0$. Since $R_1(t+1)$, $R_2(t+1)$ are not unique, by (A.38)(A.39), it requires $\lim_{t\to\infty}\|\hat{f}(t) - \hat{f}(t+1)\|^2_{R_1(t+1)} = 0$ and $\lim_{t\to\infty}\|2\Lambda(t+1) - 2\Lambda(t)\|^2_{R_2(t+1)} = 0$ should hold for all possible $R_1(t+1)$, $R_2(t+1)$. Therefore, $\lim_{t\to\infty}(\hat{f}(t) - \hat{f}(t+1)) = \mathbf{0}_{N\times d}$ and $\lim_{t\to\infty}(2\Lambda(t+1) - 2\Lambda(t)) = \mathbf{0}_{N\times d}$ should hold. $(\hat{f}(t), \Lambda(t))$ converges to the stationary point $(\hat{f}^s, \Lambda^s)$. Now show that the stationary point $(\hat{f}^s, \Lambda^s)$ is the optimal point $(\hat{f}^*, \Lambda^*)$.

Take the limit of both sides of (2.42)(2.43) yield:

$$(I + W(t+1)(D+A)\tilde{D}(t)^{-1}) \cdot (\nabla\hat{O}(\hat{f}^s, D_{all}) + 2\Lambda^s) = \mathbf{0}_{N\times d}\ ; \tag{A.42}$$

$$(D-A)\hat{f}^s = \mathbf{0}_{N\times d}\ . \tag{A.43}$$

Since $I + W(t+1)(D+A)\tilde{D}(t)^{-1} \succ \mathbf{0}_{N\times N}$, to satisfy (A.42), $\nabla\hat{O}(\hat{f}^s, D_{all}) + 2\Lambda^s = \mathbf{0}_{N\times d}$ must hold. Compare with (2.22)(2.23) in Lemma 1 and observe that $(\hat{f}^s, \Lambda^s)$ satisfies the optimality condition and is thus the optimal point. Therefore, $(\hat{f}(t), \Lambda(t))$ converges to $(\hat{f}^*, \Lambda^*)$.

A.5 Proof of Theorem 4

In the following proof, use the uppercase letters and lowercase letters to denote random variables and the corresponding realizations.

Since the modified ADMM is randomized, denote $F_i(t)$ as the random variable of the result that node i broadcasts in t-th iteration, of which the realization is $f_i(t)$. Define $F(t) = \{F_i(t)\}_{i=1}^N$ whose realization is $\{f_i(t)\}_{i=1}^N$.

Let $\mathscr{F}_{F(0:t)}(\cdot)$ be the joint probability distribution of $F(0:t) = \{F(r)\}_{r=0}^t$, and $\mathscr{F}_{F(t)}(\cdot)$ be the distribution of $F(t)$, by chain rule:

$$\begin{aligned}\mathscr{F}_{F(0:T)}(\{f(r)\}_{r=0}^T) &= \mathscr{F}_{F(0:T-1)}(\{f(r)\}_{r=0}^{T-1}) \cdot \mathscr{F}_{F(T)}(f(T)|\{f(r)\}_{r=0}^{T-1}) = \cdots \\ &= \mathscr{F}_{F(0)}(f(0)) \cdot \prod_{t=1}^{T} \mathscr{F}_{F(t)}(f(t)|\{f(r)\}_{r=0}^{t-1})\end{aligned}$$

For two neighboring datasets D_{all} and $\hat{D}_{all}$ of the network, the ratio of joint probabilities is given by:

$$\frac{\mathscr{F}_{F(0:T)}(\{f(r)\}_{r=0}^T|D_{all})}{\mathscr{F}_{F(0:T)}(\{f(r)\}_{r=0}^T|\hat{D}_{all})} = \frac{\mathscr{F}_{F(0)}(f(0)|D_{all})}{\mathscr{F}_{F(0)}(f(0)|\hat{D}_{all})} \cdot \prod_{t=1}^{T} \frac{\mathscr{F}_{F(t)}(f(t)|\{f(r)\}_{r=0}^{t-1}, D_{all})}{\mathscr{F}_{F(t)}(f(t)|\{f(r)\}_{r=0}^{t-1}, \hat{D}_{all})} \tag{A.44}$$

Since $f_i(0)$ is randomly selected for all i, which is independent of dataset, there is $\mathscr{F}_{F(0)}(f(0)|D_{all}) = \mathscr{F}_{F(0)}(f(0)|\hat{D}_{all})$.

First only consider t-th iteration, since the primal variable is updated according to (2.47), by KKT optimality condition, $\nabla_{f_i} L_i^{priv}(t)|_{f_i=f_i(t)} = 0$, implies:

$$\begin{aligned}\epsilon_i(t) = &-\frac{1}{2\eta_i(t)V_i}\frac{C}{B_i}\sum_{n=1}^{B_i} y_i^n \mathscr{L}'(y_i^n f_i(t)^T x_i^n) x_i^n - \frac{1}{2\eta_i(t)V_i}(\frac{\rho}{N}\nabla R(f_i(t)) + 2\lambda_i(t-1)) \\ &-\frac{1}{2V_i}\sum_{j\in\mathscr{V}_i}(2f_i(t) - f_i(t-1) - f_j(t-1))\end{aligned} \tag{A.45}$$

Given $\{f_i(r)\}_{r=0}^{t-1}$, $F_i(t)$ and $E_i(t)$ will be bijective:

- For any $F_i(t)$ with the realization $f_i(t)$, $\exists$ an unique $E_i(t) = \epsilon_i(t)$ having the form of (A.45) such that the KKT condition holds.
- Since the Lagrangian $L_i^{priv}(t)$ is strictly convex (by Assumption 4,5), its minimizer is unique, implies that for any $E_i(t)$ with the realization $\epsilon_i(t)$, $\exists$ an unique $F_i(t) = f_i(t)$ such that the KKT condition holds.

Since each node i generates $\epsilon_i(t)$ independently, $f_i(t)$ is also independent from each other. Let $\mathscr{F}_{F_i(t)}(\cdot)$ be the distribution of $F_i(t)$, there is:

$$\frac{\mathscr{F}_{F(t)}(f(t)|\{f(r)\}_{r=0}^{t-1}, D_{all})}{\mathscr{F}_{F(t)}(f(t)|\{f(r)\}_{r=0}^{t-1}, \hat{D}_{all})} = \prod_{v=1}^{N} \frac{\mathscr{F}_{F_v(t)}(f_v(t)|\{f_v(r)\}_{r=0}^{t-1}, D_v)}{\mathscr{F}_{F_v(t)}(f_v(t)|\{f_v(r)\}_{r=0}^{t-1}, \hat{D}_v)} = \frac{\mathscr{F}_{F_i(t)}(f_i(t)|\{f_i(r)\}_{r=0}^{t-1}, D_i)}{\mathscr{F}_{F_i(t)}(f_i(t)|\{f_i(r)\}_{r=0}^{t-1}, \hat{D}_i)} \tag{A.46}$$

Since two neighboring datasets D_{all} and $\hat{D}_{all}$ only have at most one data point that is different, the second equality holds is because of the fact that this different data point could only be possessed by one node, say node i. Then there is $D_j = \hat{D}_j$ for $j \neq i$.

Given $\{f_i(r)\}_{r=0}^{t-1}$, let $g_t(\cdot, D_i) : \mathbb{R}^d \to \mathbb{R}^d$ denote the one-to-one mapping from $E_i(t)$ to $F_i(t)$ using dataset D_i. Let $\mathscr{F}_{E_i(t)}(\cdot)$ be the probability density of $E_i(t)$, by Jacobian transformation, there is[1]:

$$\mathscr{F}_{F_i(t)}(f_i(t)|D_i) = \mathscr{F}_{E_i(t)}(g_t^{-1}(f_i(t), D_i)) \cdot |\det(\mathbf{J}(g_t^{-1}(f_i(t), D_i)))| \tag{A.47}$$

where $g_t^{-1}(f_i(t), D_i)$ is the mapping from $F_i(t)$ to $E_i(t)$ using data D_i as shown in (A.45) and $\mathbf{J}(g_t^{-1}(f_i(t), D_i))$ is the Jacobian matrix of it.

Without loss of generality, let D_i and $\hat{D}_i$ be only different in the first data point, say (x_i^1, y_i^1) and $(\hat{x}_i^1, \hat{y}_i^1)$ respectively. Then by (A.46)(A.47), (A.44) yields:

$$\frac{\mathscr{F}_{F(0:T)}(\{f(r)\}_{r=0}^{T}|D_{all})}{\mathscr{F}_{F(0:T)}(\{f(r)\}_{r=0}^{T}|\hat{D}_{all})} = \prod_{t=1}^{T} \frac{\mathscr{F}_{E_i(t)}(g_t^{-1}(f_i(t), D_i))}{\mathscr{F}_{E_i(t)}(g_t^{-1}(f_i(t), \hat{D}_i))} \cdot \prod_{t=1}^{T} \frac{|\det(\mathbf{J}(g_t^{-1}(f_i(t), D_i)))|}{|\det(\mathbf{J}(g_t^{-1}(f_i(t), \hat{D}_i)))|} \tag{A.48}$$

[1] We believe that there is a critical mistake in [147] and the original paper [24] where the objective perturbation method was proposed. A wrong mapping is used in both work:

$$\mathscr{F}_{F_i(t)}(f_i(t)|D_i) = \mathscr{F}_{E_i(t)}(g_t^{-1}(f_i(t), D_i)) \cdot |\det(\mathbf{J}(g_t^{-1}(f_i(t), D_i)))|^{-1}$$

Consider the first part, $E_i(t) \sim \exp\{-\alpha_i(t)\|\epsilon\|\}$, let $\hat{\epsilon}_i(t) = g_t^{-1}(f_i(t), \hat{D}_i)$ and $\epsilon_i(t) = g_t^{-1}(f_i(t), D_i)$

$$\prod_{t=1}^{T} \frac{\mathscr{F}_{E_i(t)}(g_t^{-1}(f_i(t), D_i))}{\mathscr{F}_{E_i(t)}(g_t^{-1}(f_i(t), \hat{D}_i))} = \prod_{t=1}^{T} \exp(\alpha_i(t)(\|\hat{\epsilon}_i(t)\| - \|\epsilon_i(t)\|)) \leq \exp(\sum_{t=1}^{T} \alpha_i(t)\|\hat{\epsilon}_i(t) - \epsilon_i(t)\|) \quad \text{(A.49)}$$

By (A.45), Assumptions 4 and the facts that $\|x_i^n\|_2 \leq 1$ (pre-normalization), $y_i^n \in \{+1, -1\}$.

$$\|\hat{\epsilon}_i(t) - \epsilon_i(t)\| = \frac{1}{2\eta_i(t)V_i}\frac{C}{B_i} \cdot \|y_i^1 \mathscr{L}'(y_i^1 f_i(t)^T x_i^1)x_i^1 - \hat{y}_i^1 \mathscr{L}'(\hat{y}_i^1 f_i(t)^T \hat{x}_i^1)\hat{x}_i^1\| \leq \frac{C}{\eta_i(t)V_i B_i}$$

(A.49) can be bounded:

$$\prod_{t=1}^{T} \frac{\mathscr{F}_{E_i(t)}(g_t^{-1}(f_i(t), D_i))}{\mathscr{F}_{E_i(t)}(g_t^{-1}(f_i(t), \hat{D}_i))} \leq \exp(\sum_{t=1}^{T} \frac{C\alpha_i(t)}{\eta_i(t)V_i B_i}) \quad \text{(A.50)}$$

Consider the second part, the Jacobian matrix $\mathbf{J}(g_t^{-1}(f_i(t), D_i))$ is:

$$\mathbf{J}(g_t^{-1}(f_i(t), D_i)) = -\frac{1}{2\eta_i(t)V_i}\frac{C}{B_i}\sum_{n=1}^{B_i} \mathscr{L}''(y_i^n f_i(t)^T x_i^n)x_i^n(x_i^n)^T - \frac{1}{2\eta_i(t)V_i}\frac{\rho}{N}\nabla^2 R(f_i(t)) - \mathbf{I}_d$$

Let $G(t) = \frac{C}{2\eta_i(t)V_i B_i}(\mathscr{L}''(\hat{y}_i^1 f_i(t)^T \hat{x}_i^1)\hat{x}_i^1(\hat{x}_i^1)^T - \mathscr{L}''(y_i^1 f_i(t)^T x_i^1)x_i^1(x_i^1)^T)$ and $H(t) = -\mathbf{J}(g_t^{-1}(f_i(t), D_i))$, there is:

$$\frac{|\det(\mathbf{J}(g_t^{-1}(f_i(t), D_i)))|}{|\det(\mathbf{J}(g_t^{-1}(f_i(t), \hat{D}_i)))|} = \frac{|\det(H(t))|}{|\det(H(t) + G(t))|}$$
$$= \frac{1}{|\det(I + H(t)^{-1}G(t))|} = \frac{1}{|\prod_{j=1}^{r}(1 + \lambda_j(H(t)^{-1}G(t)))|}$$

where $\lambda_j(H(t)^{-1}G(t))$ denotes the j-th largest eigenvalue of $H(t)^{-1}G(t)$. Since $G(t)$ has rank at most 2, implies $H(t)^{-1}G(t)$ also has rank at most 2.

Because θ is determined such that $2c_1 < \frac{B_i}{C}(\frac{\rho}{N} + 2\theta V_i)$, and $\theta \leq \eta_i(t)$ holds for all node i and iteration t, which implies:

$$\frac{c_1}{\frac{B_i}{C}(\frac{\rho}{N} + 2\eta_i(t)V_i)} < \frac{1}{2} \quad \text{(A.51)}$$

By Assumptions 4 and 5, the eigenvalue of $H(t)$ and $G(t)$ satisfy:

$$\lambda_j(H(t)) \geq \frac{\rho}{2\eta_i(t)V_iN} + 1 > 0$$

$$-\frac{Cc_1}{2\eta_i(t)V_iB_i} \leq \lambda_j(G(t)) \leq \frac{Cc_1}{2\eta_i(t)V_iB_i}$$

Implies:

$$-\frac{c_1}{\frac{B_i}{C}(\frac{\rho}{N} + 2\eta_i(t)V_i)} \leq \lambda_j(H(t)^{-1}G(t)) \leq \frac{c_1}{\frac{B_i}{C}(\frac{\rho}{N} + 2\eta_i(t)V_i)}$$

By (A.51):

$$-\frac{1}{2} \leq \lambda_j(H(t)^{-1}G(t)) \leq \frac{1}{2}$$

Since $\lambda_{\min}(H(t)^{-1}G(t)) > -1$, there is:

$$\frac{1}{|1+\lambda_{\max}(H(t)^{-1}G(t))|^2} \leq \frac{1}{|\det(I+H(t)^{-1}G(t))|} \leq \frac{1}{|1+\lambda_{\min}(H(t)^{-1}G(t))|^2}$$

Therefore,

$$\prod_{t=1}^{T}\frac{|\det(\mathbf{J}(g_t^{-1}(f_i(t),D_i)))|}{|\det(\mathbf{J}(g_t^{-1}(f_i(t),\hat{D}_i)))|} \leq \prod_{t=1}^{T}\frac{1}{|1-\frac{c_1}{\frac{B_i}{C}(\frac{\rho}{N}+2\eta_i(t)V_i)}|^2} = \exp(-\sum_{t=1}^{T}2\ln(1-\frac{c_1}{\frac{B_i}{C}(\frac{\rho}{N}+2\eta_i(t)V_i)})) \tag{A.52}$$

Since for any real number $x \in [0, 0.5]$, $-\ln(1-x) < 1.4x$. By condition (A.51), (A.52) can be bounded with a simper expression:

$$\prod_{t=1}^{T}\frac{|\det(\mathbf{J}(g_t^{-1}(f_i(t),D_i)))|}{|\det(\mathbf{J}(g_t^{-1}(f_i(t),\hat{D}_i)))|} \leq \exp(\sum_{t=1}^{T}\frac{2.8c_1}{\frac{B_i}{C}(\frac{\rho}{N}+2\eta_i(t)V_i)}) \leq \exp(\sum_{t=1}^{T}\frac{1.4Cc_1}{\eta_i(t)V_iB_i}) \tag{A.53}$$

Combine (A.50)(A.53), (A.48) can be bounded:

$$\frac{\mathscr{F}_{F(0:T)}(\{f(r)\}_{r=0}^{T}|D_{all})}{\mathscr{F}_{F(0:T)}(\{f(r)\}_{r=0}^{T}|\hat{D}_{all})} \leq \exp(\sum_{t=1}^{T}(\frac{1.4Cc_1}{\eta_i(t)V_iB_i} + \frac{C\alpha_i(t)}{\eta_i(t)V_iB_i})) = \exp(\sum_{t=1}^{T}\frac{C}{\eta_i(t)V_iB_i}(1.4c_1+\alpha_i(t)))$$

Therefore, the total privacy loss during T iterations can be bounded by any β:

$$\beta \geq \max_{i \in \mathcal{N}} \{ \sum_{t=1}^{T} \frac{C}{\eta_i(t) V_i B_i} (1.4 c_1 + \alpha_i(t)) \}$$

A.6 Proof of Lemma 2

Consider the private MR-ADMM up to $2k$-th iteration. In $(2k-1)$-th iteration, the primal variable is updated via (2.47), By KKT condition:

$$\begin{aligned} \nabla O(f_i(2k-1), D_i) + \epsilon_i(2k-1) &= -2\lambda_i(2k-2) \\ -\eta_i(2k-1) \sum_{j \in \mathcal{V}_i} (2 f_i(2k-1) - f_i(2k-2) - f_j(2k-2)) & \end{aligned} \tag{A.54}$$

Given $\{f_i(t)\}_{i=1}^N$ for $t \leq 2k-2$, $\{\lambda_i(2k-2)\}_{i=1}^N$ are also given. RHS of (A.54) can be calculated completely after releasing $\{f_i(k-1)\}_{i=1}^N$, i.e., the information of $\nabla O(f_i(2k-1), D_i) + \epsilon_i(2k-1)$ is completely released during $(2k-1)$-th iteration. Suppose the private MR-ADMM satisfies β_{2k-1}-differential privacy during $(2k-1)$ iterations, then in $(2k)$-th iterations, by (2.48):

$$\begin{aligned} f_i(2k) \quad = \quad & f_i(2k-1) - \frac{1}{2\eta V_i + \gamma} \{ \nabla O(f_i(2k-1), D_i) + \epsilon_i(2k-1) + 2\lambda_i(2k-1) \\ & + \eta_i(2k-1) \sum_{j \in \mathcal{V}_i} (f_i(2k-1) - f_j(2k-1)) \} \end{aligned}$$

which is a deterministic mapping taking the outputs from $(2k-1)$-th iteration as input. Because the differential privacy is immune to post-processing [37], releasing $\{f_i(2k)\}_{i=1}^N$ doesn't increase the privacy loss, i.e., the total privacy loss up to $(2k)$-th iteration can still be bounded by β_{2k-1}.

A.7 Proof of Theorem 5

Use the uppercase letters X and lowercase letters x to denote random variables and the corresponding realizations, and use $\mathscr{F}_X(\cdot)$ to denote its probability distribution.

For two neighboring datasets D_{all} and $\hat{D}_{all}$ of the network, by Lemma 2, the total privacy loss is

only contributed by odd iterations. Thus, the ratio of joint probabilities (privacy loss) is given by:

$$\frac{\mathscr{F}_{F(0:2K)}(\{f(r)\}_{r=0}^{2}K|D_{all})}{\mathscr{F}_{F(0:2K)}(\{f(r)\}_{r=0}^{2}K|\hat{D}_{all})} = \frac{\mathscr{F}_{F(0)}(f(0)|D_{all})}{\mathscr{F}_{F(0)}(f(0)|\hat{D}_{all})}\prod_{k=1}^{K}\frac{\mathscr{F}_{F(2k-1)}(f(2k-1)|\{f(r)\}_{r=0}^{2k-2},D_{all})}{\mathscr{F}_{F(2k-1)}(f(2k-1)|\{f(r)\}_{r=0}^{2t-2},\hat{D}_{all})} \quad \text{(A.55)}$$

Since $f_i(0)$ is randomly selected for all i, which is independent of dataset, there is $\mathscr{F}_{F(0)}(f(0)|D_{all}) = \mathscr{F}_{F(0)}(f(0)|\hat{D}_{all})$. First only consider $(2k-1)$-th iteration, since the primal variable is updated according to (2.47), by KKT optimality condition:

$$\begin{aligned}\epsilon_i(2k-1) &= -\nabla O(f_i(2k-1),D_i) - 2\lambda_i(2k-2)\\ &\quad -\eta_i(2k-1)\sum_{j\in\mathscr{V}_i}(2f_i(2k-1) - f_i(2k-2) - f_j(2k-2))\end{aligned} \quad \text{(A.56)}$$

Given $\{f(r)\}_{r=0}^{2k-2}$, $F_i(2k-1)$ and $E_i(2k-1)$ will be bijective $\forall i$, there is:

$$\begin{aligned}\frac{\mathscr{F}_{F(2k-1)}(f(2k-1)|\{f(r)\}_{r=0}^{2k-2},D_{all})}{\mathscr{F}_{F(2k-1)}(f(2k-1)|\{f(r)\}_{r=0}^{2k-2},\hat{D}_{all})} &= \prod_{v=1}^{N}\frac{\mathscr{F}_{F_v(2k-1)}(f_v(2k-1)|\{f_v(r)\}_{r=0}^{2k-2},D_v)}{\mathscr{F}_{F_v(2k-1)}(f_v(2k-1)|\{f_v(r)\}_{r=0}^{2k-2},\hat{D}_v)}\\ &= \frac{\mathscr{F}_{F_i(2k-1)}(f_i(2k-1)|\{f_i(r)\}_{r=0}^{2k-2},D_i)}{\mathscr{F}_{F_i(2k-1)}(f_i(2k-1)|\{f_i(r)\}_{r=0}^{2k-2},\hat{D}_i)}\end{aligned} \quad \text{(A.57)}$$

Since two neighboring datasets D_{all} and $\hat{D}_{all}$ only have at most one data point that is different, the second equality holds is because of the fact that this different data point could only be possessed by one node, say node i. Then there is $D_j = \hat{D}_j$ for $j \neq i$.

Given $\{f(r)\}_{r=0}^{2k-2}$, let $g_k(\cdot, D_i) : \mathbb{R}^d \to \mathbb{R}^d$ denote the one-to-one mapping from $E_i(2k-1)$ to $F_i(2k-1)$ using dataset D_i. By Jacobian transformation, there is $\mathscr{F}_{F_i(2k-1)}(f_i(2k-1)|D_i) = \mathscr{F}_{E_i(2k-1)}(g_k^{-1}(f_i(2k-1),D_i))\cdot|\det(\mathbf{J}(g_k^{-1}(f_i(2k-1),D_i)))|$, where $g_k^{-1}(f_i(2k-1),D_i)$ is the mapping from $F_i(2k-1)$ to $E_i(2k-1)$ using data D_i as shown in (A.56) and $\mathbf{J}(g_k^{-1}(f_i(2k-1),D_i))$ is the Jacobian matrix of it. Then (A.55) yields:

$$\frac{\mathscr{F}_{F(0:2K)}(\{f(r)\}_{r=0}^{2K}|D_{all})}{\mathscr{F}_{F(0:2K)}(\{f(r)\}_{r=0}^{2K}|\hat{D}_{all})} = \prod_{k=1}^{K}\frac{\mathscr{F}_{E_i(2k-1)}(g_k^{-1}(f_i(2k-1),D_i))}{\mathscr{F}_{E_i(2k-1)}(g_k^{-1}(f_i(2k-1),\hat{D}_i))}\prod_{k=1}^{K}\frac{|\det(\mathbf{J}(g_k^{-1}(f_i(2k-1),D_i)))|}{|\det(\mathbf{J}(g_k^{-1}(f_i(2k-1),\hat{D}_i)))|} \quad \text{(A.58)}$$

Consider the first part, $E_i(2k-1) \sim \exp\{-\alpha_i(k)\|\epsilon\|\}$, let $\hat{\epsilon}_i(2k-1) = g_k^{-1}(f_i(2k-1),\hat{D}_i)$ and $\epsilon_i(2k-1) =$

$g_k^{-1}(f_i(2k-1), D_i)$

$$\begin{aligned}\prod_{k=1}^{K} \frac{\mathscr{F}_{E_i(2k-1)}(g_k^{-1}(f_i(2k-1), D_i))}{\mathscr{F}_{E_i(2k-1)}(g_k^{-1}(f_i(2k-1), \hat{D}_i))} &= \prod_{k=1}^{K} \exp(\alpha_i(k)(\|\hat{\epsilon}_i(2k-1)\| - \|\epsilon_i(2k-1)\|)) \\ &\leq \exp(\sum_{k=1}^{K} \alpha_i(k)\|\hat{\epsilon}_i(2k-1) - \epsilon_i(2k-1)\|) \end{aligned} \tag{A.59}$$

Without loss of generality, let D_i and $\hat{D}_i$ be only different in the first data point, say (x_i^1, y_i^1) and $(\hat{x}_i^1, \hat{y}_i^1)$ respectively. By (A.56), Assumptions 4 and the facts that $\|x_i^n\|_2 \leq 1$ (pre-normalization), $y_i^n \in \{+1, -1\}$.

$$\|\hat{\epsilon}_i(2k-1) - \epsilon_i(2k-1)\| = \|\nabla O(f_i(2k-1), \hat{D}_i) - \nabla O(f_i(2k-1), D_i)\| \leq \frac{2C}{B_i} \tag{A.60}$$

(A.59) can be bounded:

$$\prod_{k=1}^{K} \frac{\mathscr{F}_{E_i(2k-1)}(g_k^{-1}(f_i(2k-1), D_i))}{\mathscr{F}_{E_i(2k-1)}(g_k^{-1}(f_i(2k-1), \hat{D}_i))} \leq \exp(\sum_{k=1}^{K} \frac{2C\alpha_i(k)}{B_i}) \tag{A.61}$$

Consider the second part, the Jacobian matrix $\mathbf{J}(g_k^{-1}(f_i(2k-1), D_i))$ is:

$$\begin{aligned}\mathbf{J}(g_k^{-1}(f_i(2k-1), D_i)) &= -\frac{C}{B_i} \sum_{n=1}^{B_i} \mathscr{L}''(y_i^n f_i(2k-1)^T x_i^n) x_i^n (x_i^n)^T \\ &\quad -\frac{\rho}{N} \nabla^2 R(f_i(2k-1)) - 2\eta_i(2k-1) V_i \mathbf{I}_d \end{aligned}$$

Define

$$\begin{aligned} G(k) &= \frac{C}{B_i}(\mathscr{L}''(\hat{y}_i^1 f_i(2k-1)^T \hat{x}_i^1)\hat{x}_i^1(\hat{x}_i^1)^T - \mathscr{L}''(y_i^1 f_i(2k-1)^T x_i^1) x_i^1 (x_i^1)^T) \; ; \\ H(k) &= -\mathbf{J}(g_k^{-1}(f_i(2k-1), D_i)) \; . \end{aligned}$$

There is:

$$\begin{aligned}\frac{|\det(\mathbf{J}(g_k^{-1}(f_i(2k-1),D_i)))|}{|\det(\mathbf{J}(g_k^{-1}(f_i(2k-1),\hat{D}_i)))|} &= \frac{|\det(H(k))|}{|\det(H(k)+G(k))|} = \frac{1}{|\det(I+H(k)^{-1}G(k))|} \\ &= \frac{1}{|\prod_{j=1}^{r}(1+\lambda_j(H(k)^{-1}G(k)))|} \end{aligned} \tag{A.62}$$

where $\lambda_j(H(k)^{-1}G(k))$ denotes the j-th largest eigenvalue of $H(k)^{-1}G(k)$. Since $G(k)$ has rank at most 2, $H(k)^{-1}G(k)$ also has rank at most 2. By Assumptions 4 and 5, the eigenvalue of $H(k)$ and $G(k)$ satisfy

$$\begin{aligned}&\lambda_j(H(k)) \geq \frac{\rho}{N} + 2\eta_i(2k-1)V_i > 0\ ; \\ &-\frac{Cc_1}{B_i} \leq \lambda_j(G(k)) \leq \frac{Cc_1}{B_i}\ .\end{aligned}$$

Implies

$$-\frac{c_1}{\frac{B_i}{C}(\frac{\rho}{N}+2\eta_i(2k-1)V_i)} \leq \lambda_j(H(k)^{-1}G(k)) \leq \frac{c_1}{\frac{B_i}{C}(\frac{\rho}{N}+2\eta_i(2k-1)V_i)}\ .$$

Since $2c_1 < \frac{B_i}{C}(\frac{\rho}{N}+2\eta_i(1)V_i)$ and $\eta_i(2k-1) \leq \eta_i(2k+1)$ for all k, $2c_1 < \frac{B_i}{C}(\frac{\rho}{N}+2\eta_i(2k-1)V_i)$ holds. It implies the following,

$$-\frac{1}{2} \leq \lambda_j(H(k)^{-1}G(k)) \leq \frac{1}{2}.$$

Since $\lambda_{\min}(H(k)^{-1}G(k)) > -1$, there is

$$\frac{1}{|1+\lambda_{\max}(H(k)^{-1}G(k))|^2} \leq \frac{1}{|\det(I+H(k)^{-1}G(k))|} \leq \frac{1}{|1+\lambda_{\min}(H(k)^{-1}G(k))|^2}\ .$$

Therefore,

$$\prod_{k=1}^{K}\frac{|\det(\mathbf{J}(g_k^{-1}(f_i(2k-1),D_i)))|}{|\det(\mathbf{J}(g_k^{-1}(f_i(2k-1),\hat{D}_i)))|} \leq \prod_{k=1}^{K}\frac{1}{|1-\frac{c_1}{\frac{B_i}{C}(\frac{\rho}{N}+2\eta_i(2k-1)V_i)}|^2}$$
$$= \exp(-\sum_{k=1}^{K}2\ln(1-\frac{c_1}{\frac{B_i}{C}(\frac{\rho}{N}+2\eta_i(2k-1)V_i)})) \quad (A.63)$$

Since for any real number $x \in [0, 0.5]$, $-\ln(1-x) < 1.4x$. (A.63) can be bounded with a simper expression:

$$\prod_{k=1}^{K}\frac{|\det(\mathbf{J}(g_k^{-1}(f_i(2k-1),D_i)))|}{|\det(\mathbf{J}(g_k^{-1}(f_i(2k-1),\hat{D}_i)))|} \leq \exp(\sum_{k=1}^{K}\frac{2.8c_1}{\frac{B_i}{C}(\frac{\rho}{N}+2\eta_i(2k-1)V_i)}). \quad (A.64)$$

Combine (A.61)(A.64), (A.58) can be bounded:

$$\frac{\mathscr{F}_{F(0:2K)}(\{f(r)\}_{r=0}^{2K}|D_{all})}{\mathscr{F}_{F(0:2K)}(\{f(r)\}_{r=0}^{2K}|\hat{D}_{all})} \leq \exp(\sum_{k=1}^{K}\frac{2C}{B_i}(\frac{1.4c_1}{(\frac{\rho}{N}+2\eta_i(2k-1)V_i)}+\alpha_i(k))) \ . \quad (A.65)$$

Therefore, the total privacy loss during T iterations can be bounded by any β:

$$\beta \geq \max_{i\in\mathscr{N}}\{\sum_{k=1}^{K}\frac{2C}{B_i}(\frac{1.4c_1}{(\frac{\rho}{N}+2\eta_i(2k-1)V_i)}+\alpha_i(k))\} \ .$$

A.8 Proof of Theorem 6

Let $\widetilde{O}(f) = C\mathcal{L}(f) + \frac{\rho}{2N}\|f\|^2$ and $\widetilde{f_i} = \text{argmin}_f \widetilde{O}(f)$. Let $f_i^{opt} = \text{argmin}_f O(f, D_i)$ be node i's classifier trained with its own data.

$$\mathcal{L}(f_c^*) = \mathcal{L}(f_{ref}) + (\frac{\widetilde{O}(f_c^*)}{C} - \frac{\widetilde{O}(\widetilde{f_i})}{C}) + (\frac{\rho}{2NC}\|f_{ref}\|^2 - \frac{\rho}{2NC}\|f_c^*\|^2) + \Big(\frac{\widetilde{O}(\widetilde{f_i})}{C} - \frac{\widetilde{O}(f_{ref})}{C}\Big)$$

By [126], $\widetilde{O}(f_c^*) - \widetilde{O}(\widetilde{f_i}) \leq (1+a)(O(f_c^*, D_i) - O(f_i^{opt}, D_i)) + \mathcal{O}(\frac{C^2 N \log(1/\delta)}{\rho B_i})$ holds $\forall a > 0$ with probability $1-\delta$, where $\mathcal{O}$ is big-$\mathcal{O}$ notation.

Since f_c^* is the centralized classifier trained with samples from all nodes, we assume the difference of empirical loss under two classifiers f_c^* and f_i^{opt} is bounded by $\nu > 0$, i.e., $O(f_c^*, D_i) -$

$O(f_i^{opt}, D_i) \leq \frac{\rho}{2N}(\|f_c^*\|^2 - \|f_i^{opt}\|^2) + C\nu$. Moreover, $\widetilde{O}(\widetilde{f_i}) \leq \widetilde{O}(f_{ref})$.

$$\begin{aligned}\mathcal{L}(f_c^*) \quad \leq \quad & \mathcal{L}(f_{ref}) + \mathcal{O}(\frac{CN\log(1/\delta)}{\rho B_i}) + (1+a)(\frac{\rho}{2NC}\|f_c^*\|^2 - \frac{\rho}{2NC}\|f_i^{opt}\|^2 + \nu) \\ & + (\frac{\rho}{2NC}\|f_{ref}\|^2 - \frac{\rho}{2NC}\|f_c^*\|^2)\end{aligned}$$

We assume ν is relatively small as compared to other terms. If choosing $a > 0$ to be a sufficient small number such that $a\|f_c^*\|^2 - (1+a)\|f_i^{opt}\|^2 \leq 0$ and choosing ρ such that $\frac{\rho}{2NC}\|f_{ref}\|^2 \leq \frac{\tau - \Delta_i(k)}{2}$, e.g., $\rho \leq \frac{NC(\tau - \Delta_i(k))}{\|f_{ref}\|^2}$, and if B_i also satisfies $\mathcal{O}(\frac{CN\log(1/\delta)}{\rho B_i}) \leq \frac{\tau - \Delta_i(k)}{2}$, i.e.,

$$B_i \geq w \max_k \{\frac{CN\log(1/\delta)}{\rho(\tau - \Delta_i(k))}\} \geq w \max_k \{\frac{\|f_{ref}\|^2 \log(1/\delta)}{(\tau - \Delta_i(k))^2}\}$$

for some constant w, then the following holds with probability $1 - \delta$.

$$\mathcal{L}(f_c^*) \quad \leq \quad \mathcal{L}(f_{ref}) + \tau - \Delta_i(k)$$

Since $\mathcal{L}(f_i^{non}(2k-1)) \leq \mathcal{L}(f_c^*) + \Delta_i(k)$, it implies that $\mathcal{L}(f_i^{non}(2k-1)) \leq \mathcal{L}(f_{ref}) + \tau$ holds with probability $1 - \delta$.

A.9 proof of Theorem 7

Let $\widetilde{O}(f) = C\mathcal{L}(f) + \frac{\rho}{2N}\|f\|^2$ and $\widetilde{f_i} = \text{argmin}_f \widetilde{O}(f)$. Let $f_i^{opt} = \text{argmin}_f O(f, D_i)$ be node i's classifier trained with its own data. Let $f_i^{privOpt} = \text{argmin}_f O^{priv}(f, D_i; \epsilon) = O(f, D_i) + \epsilon^T f$ and $\widetilde{O}^{priv}(f; \epsilon) = \widetilde{O}(f) + \epsilon^T f$.

$$\begin{aligned}\mathcal{L}(f_{new}^*) \quad = \quad & \mathcal{L}(f_{ref}) + (\frac{\widetilde{O}(f_{new}^*)}{C} - \frac{\widetilde{O}(f_i^{privOpt})}{C}) + (\frac{\widetilde{O}(f_i^{privOpt})}{C} - \frac{\widetilde{O}(\widetilde{f_i})}{C}) \\ & + (\frac{\rho}{2NC}\|f_{ref}\|^2 - \frac{\rho}{2NC}\|f_{new}^*\|^2) + (\frac{\widetilde{O}(\widetilde{f_i})}{C} - \frac{\widetilde{O}(f_{ref})}{C})\end{aligned}$$

For the new optimization problem, f_{new}^* is centralized classifier trained with samples from all nodes while $f_i^{privOpt}$ is the classifier trained with samples from node i. We assume the difference of empirical loss under two classifiers f_{new}^* and $f_i^{privOpt}$ can be bounded by $\nu > 0$, i.e., $\widetilde{O}(f_{new}^*) -$

$\widetilde{O}(f_i^{privOpt}) \leq \frac{\rho}{2N}(\|f_{new}^*\|^2 - \|f_i^{privOpt}\|^2) + C\nu$.

By [126], $\widetilde{O}(f_i^{privOpt}) - \widetilde{O}(\widetilde{f_i}) \leq (1+a)(O(f_i^{privOpt}, D_i) - O(f_i^{opt}, D_i)) + \mathcal{O}(\frac{C^2 N \log(1/\delta)}{\rho B_i})$ holds $\forall a > 0$ with probability $1-\delta$. By Lemma 12, $O(f_i^{privOpt}, D_i) - O(f_i^{opt}, D_i) \leq \frac{Nd^2}{\rho(\alpha_i(k))^2}(\log(d/\delta))^2$ holds with probability $1-\delta$. Therefore, $\widetilde{O}(f_i^{privOpt}) - \widetilde{O}(\widetilde{f_i}) \leq (1+a)(\frac{Nd^2}{\rho(\alpha_i(k))^2}(\log(d/\delta))^2) + \mathcal{O}(\frac{C^2 N \log(1/\delta)}{\rho B_i})$ holds $\forall a > 0$ with probability $1-2\delta$.

Since $\widetilde{f_i} = \text{argmin}_f \widetilde{O}(f)$, implying $\widetilde{O}(\widetilde{f_i}) \leq \widetilde{O}(f_{ref})$. The following holds $\forall a > 0$ with probability $1-2\delta$,

$$\begin{aligned}\mathcal{L}(f_{new}^*) \quad \leq \quad & \mathcal{L}(f_{ref}) + \nu + \mathcal{O}(\frac{CN\log(1/\delta)}{\rho B_i}) + (1+a)\frac{Nd^2}{C\rho(\alpha_i(k))^2}(\log(d/\delta))^2 \\ & + (\frac{\rho}{2NC}\|f_{ref}\|^2 - \frac{\rho}{2NC}\|f_i^{privOpt}\|^2)\end{aligned}$$

We assume ν is relatively small as compared to other terms. If choosing ρ such that $\frac{\rho}{2NC}\|f_{ref}\|^2 \leq \frac{1}{2}(\tau - \Delta_i^{new}(k))$, i.e., $\rho \leq \frac{NC(\tau - \Delta_i^{new}(k))}{\|f_{ref}\|^2}$, and if B_i also satisfies $((1+a)\frac{Nd^2}{C(\alpha_i(k))^2}(\log(d/\delta))^2 + \mathcal{O}(\frac{CN\log(1/\delta)}{B_i})) \leq \frac{\rho(\tau - \Delta_i^{new}(k))}{2}$, i.e., $B_i \geq w\frac{CN\log(1/\delta)}{\frac{\rho(\tau-\Delta_i^{new}(k))}{2} - (1+a)\frac{Nd^2}{C(\alpha_i(k))^2}(\log(d/\delta))^2}$ for some $a > 0$ and constant w. Then $\mathcal{L}(f_{new}^*) \leq \mathcal{L}(f_{ref}) + \tau - \Delta_i^{new}(k)$ holds with probability $1-2\delta$. Plug in $\rho = \frac{NC(\tau - \Delta_i^{new}(k))}{\|f_{ref}\|^2}$ and re-organize gives:

$$B_i \geq w \max_k \{\frac{CN\log(1/\delta)}{\frac{NC(\tau-\Delta_i^{new}(k))^2}{2\|f_{ref}\|^2} - (1+a)\frac{Nd^2}{C(\alpha_i(k))^2}(\log(d/\delta))^2}\}$$

Since $\mathcal{L}(f_i^{new}(2k-1)) \leq \mathcal{L}(f_{new}^*) + \Delta_i^{new}(k)$, it implies that $\mathcal{L}(f_i^{new}(2k-1)) \leq \mathcal{L}(f_{ref}) + \tau$ holds with probability $1-2\delta$.

Lemma 12. *Let $f_i^{privOpt} = argmin_f O(f, D_i) + \epsilon^T f$ and $f_i^{opt} = argmin_f O(f, D_i)$ be outputs at iteration $2k-1$, then $O(f_i^{privOpt}, D_i) - O(f_i^{opt}, D_i) \leq \frac{Nd^2}{\rho(\alpha_i(k))^2}(\log(d/\delta))^2$ holds with probability $1-\delta$.*

Proof. There is $O(f_i^{privOpt}, D_i) \leq O(f_i^{opt}, D_i) + \epsilon^T(f_i^{opt} - f_i^{privOpt})$. By Lemma 13, since $O(f, D_i)$ and $O(f, D_i) + \epsilon^T f$ are $\frac{\rho}{N}$-strongly convex, $\|f_i^{opt} - f_i^{privOpt}\| \leq \frac{N}{\rho}\|\epsilon\|$ holds. By Lemma 14, with probability $1-\delta$, $\|\epsilon\| \leq \frac{d}{\alpha_i(k)}\log(d/\delta)$. Therefore, $O(f_i^{privOpt}, D_i) - O(f_i^{opt}, D_i) \leq \|\epsilon\|\|f_i^{opt} - f_i^{privOpt}\| \leq \frac{Nd^2}{\rho(\alpha_i(k))^2}(\log(d/\delta))^2$ holds with probability $1-\delta$. □

Lemma 13. *[24] Let $G(f)$, $g(f)$ be two vector-valued functions, which are continuous and differentiable at all points. Moreover, let $G(f)$ and $G(f)+g(f)$ be λ-strongly convex. If $f_1 = argmin_f G(f)$ and $f_2 = argmin_f G(f)+g(f)$, then $\|f_1 - f_2\| \leq \frac{1}{\lambda}\max_f \|\nabla g(f)\|$.*

Lemma 14. *[24] Let X be a random variable drawn from distribution $\Gamma(k,\theta)$, where k is an integer, then $\Pr(X < k\theta\log(k/\delta)) \geq 1-\delta$.*

APPENDIX B

Real-Time Release of Sequential Data with Differential Privacy

B.1 Proof of Propositions

The proofs are straightforward but we provide them in details for the completeness of chapter.

Finding the MMSE estimate of Z_{t+1} given $Z_t = z_t$ is equivalent to finding the mapping

$$f^* = \underset{f}{\operatorname{argmin}}\ \mathbb{E}((Z_{t+1} - f(Z_t))^2 | Z_t = z_t) = \underset{f}{\operatorname{argmin}} \int_{-\infty}^{\infty} p(z_{t+1}|z_t)(z_{t+1} - f(z_t))^2 dz_{t+1}$$

Differentiating with respect to f and equating the result to zero gives:

$$\int_{-\infty}^{\infty} p(z_{t+1}|z_t) f^*(z_t) dz_{t+1} = f^*(z_t) = \int_{-\infty}^{\infty} p(z_{t+1}|z_t) z_{t+1} dz_{t+1} = \mathbb{E}(Z_{t+1}|Z_t = z_t)$$

Therefore, the MMSE estimate of Z_{t+1} given $Z_t = z_t$ is $\mathbb{E}(Z_{t+1}|Z_t = z_t)$.

(1) Proposition 1: Gaussian AR(1) Process

(i) Since (Z_t, Z_{t+1}) is jointly Gaussian:

$$\begin{pmatrix} Z_t \\ Z_{t+1} \end{pmatrix} \sim \mathcal{N}(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_z^2 & \rho\sigma_z^2 \\ \rho\sigma_z^2 & \sigma_z^2 \end{bmatrix})$$

implies that $Z_{t+1}|Z_t \sim \mathcal{N}(\mu(1-\rho) + \rho Z_t, \sigma_z^2(1-\rho^2))$, combine with the above result implies the MMSE estimate of Z_{t+1} given $Z_t = z_t$ is $\mathbb{E}(Z_{t+1}|Z_t = z_t) = \mu(1-\rho) + \rho z_t$

The corresponding MSE is:

$$\mathbb{E}((Z_{t+1} - \mathbb{E}(Z_{t+1}|Z_t = z_t))^2|Z_t = z_t) = \text{Var}(Z_{t+1}|Z_t = z_t) = \sigma_z^2(1-\rho^2)$$

(ii) Since $Z_i \sim \mathcal{N}(\mu, \sigma_z^2)$, $N_i \sim \mathcal{N}(0, \sigma_n^2)$, there is $X_i = Z_i + N_i \sim \mathcal{N}(\mu, \sigma_z^2 + \sigma_n^2)$ and $\text{Corr}(X_i Z_{t+1}) = \rho^{t+1-i}\frac{\sigma_z}{\sqrt{\sigma_z^2+\sigma_n^2}}$.

(X_i, Z_{t+1}) is jointly Gaussian:

$$\begin{pmatrix} X_i \\ Z_{t+1} \end{pmatrix} \sim \mathcal{N}(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_z^2 + \sigma_n^2 & \rho^{t+1-i}\sigma_z^2 \\ \rho^{t+1-i}\sigma_z^2 & \sigma_z^2 \end{bmatrix})$$

implies the MMSE estimate of Z_{t+1} given $X_i = x_i$ is

$$\mathbb{E}(Z_{t+1}|X_i = x_i) = \mu(1 - \rho^{t+1-i}\frac{\sigma_z^2}{\sigma_z^2+\sigma_n^2}) + \rho^{t+1-i}\frac{\sigma_z^2}{\sigma_z^2+\sigma_n^2}x_i$$

The corresponding MSE is:

$$\sigma_z^2(1 - (\rho^{t+1-i})^2\frac{\sigma_z^2}{\sigma_z^2+\sigma_n^2})$$

(2) Proposition 2: Binomial AR(1) Process

The MMSE estimate of Z_{t+1} given $Z_t = z_t$ is $\mathbb{E}(Z_{t+1}|Z_t = z_t)$. Since the thinning is performed

independently, given $Z_t = z_t$, the probability generating function satisfies the following:

$$\begin{aligned}
G(s) &= \mathbb{E}_{Z_{t+1}|Z_t=z_t}(s^{Z_{t+1}}|Z_t = z_t) = \mathbb{E}(s^{\alpha\circ Z_t}|Z_t = z_t)\mathbb{E}(s^{\beta\circ(n-Z_t)}|Z_t = z_t) \\
&= (1-\beta+\beta s)^n(\frac{1-\alpha+\alpha s}{1-\beta+\beta s})^{z_t} \\
G'(s) &= n\beta(1-\beta+\beta s)^{n-1}(\frac{1-\alpha+\alpha s}{1-\beta+\beta s})^{z_t} + (1-\beta+\beta s)^n z_t(\frac{1-\alpha+\alpha s}{1-\beta+\beta s})^{z_t-1}\frac{\alpha}{\beta}\frac{\frac{1}{\beta}-\frac{1}{\alpha}}{(\frac{1}{\beta}-1+s)^2} \\
G''(s) &= n\beta^2(n-1)(1-\beta+\beta s)^{n-2}(\frac{1-\alpha+\alpha s}{1-\beta+\beta s})^{z_t} \\
&\quad +2n\beta(1-\beta+\beta s)^{n-1}z_t(\frac{1-\alpha+\alpha s}{1-\beta+\beta s})^{z_t-1}\frac{\alpha}{\beta}\frac{\frac{1}{\beta}-\frac{1}{\alpha}}{(\frac{1}{\beta}-1+s)^2} \\
&\quad +(1-\beta+\beta s)^n z_t((z_1-1)(\frac{1-\alpha+\alpha s}{1-\beta+\beta s})^{z_t-2}(\frac{\alpha}{\beta}\frac{\frac{1}{\beta}-\frac{1}{\alpha}}{(\frac{1}{\beta}-1+s)^2})^2 \\
&\quad +(\frac{1-\alpha+\alpha s}{1-\beta+\beta s})^{z_t-1}\frac{\alpha}{\beta}2\frac{\frac{1}{\alpha}-\frac{1}{\beta}}{(\frac{1}{\beta}-1+s)^3})
\end{aligned}$$

Since $\beta = \pi(1-\rho)$, $\alpha = \beta+\rho$, $\mathbb{E}(Z_{t+1}|Z_t = z_t) = \lim_{s\to 1} G'(s)$ and $\mathrm{Var}(Z_{t+1}|Z_t = z_t) = \lim_{s\to 1} G''(s) + G'(s) - (G'(s))^2$ gives:

$$\begin{aligned}
\mathbb{E}(Z_{t+1}|Z_t = z_t) &= \rho z_t + n\pi(1-\rho)\ ; \\
\mathrm{Var}(Z_{t+1}|Z_t = z_t) &= \rho(1-\rho)(1-2\pi)z_t + n\beta(1-\beta).
\end{aligned}$$

The corresponding MSE is:

$$\mathbb{E}((Z_{t+1} - \mathbb{E}(Z_{t+1}|Z_t = z_t))^2|Z_t = z_t) = \mathrm{Var}(Z_{t+1}|Z_t = z_t)$$

B.2 Proof of Lemma 4

Consider any $d, \hat{d} \in \mathcal{D}$, and with them the binomial mechanism outputs the same results b. Let $\bar{b} = b - \mathcal{Q}(d)$

$$\frac{\Pr(b = \mathcal{Q}(d) + noise)}{\Pr(b = \mathcal{Q}(\hat{d}) + \hat{noise})} = \frac{\mathscr{F}_N(b - \mathcal{Q}(d))}{\mathscr{F}_N(b - \mathcal{Q}(\hat{d}))} = \frac{\binom{2m}{m+b-\mathcal{Q}(d)}}{\binom{2m}{m+b-\mathcal{Q}(\hat{d})}} = \frac{\binom{2m}{m+\bar{b}}}{\binom{2m}{m+\bar{b}+\Delta\mathcal{Q}}} = \prod_{i=1}^{\Delta\mathcal{Q}} \frac{m+\bar{b}+i}{m-\bar{b}+1-i} \tag{B.1}$$

A sufficient condition for (B.1) being bounded by $\exp(\epsilon)$ is:

$$\forall i \in \{1, 2, \cdots, \Delta\mathcal{Q}\}, \quad \frac{m+\bar{b}+i}{m-\bar{b}+1-i} \leq \exp(\frac{\epsilon}{\Delta\mathcal{Q}}) \overset{\tilde{b}=\bar{b}+i}{\Longleftrightarrow} \quad \frac{m+\tilde{b}}{m-\tilde{b}+1} \leq \exp(\frac{\epsilon}{\Delta\mathcal{Q}}) \tag{B.2}$$

from (B.2), we have:

$$\bar{b} \quad \leq \quad \min_{i \in [\Delta\mathcal{Q}]} \; m + 1 - \frac{2m+1}{\exp(\frac{\epsilon}{\Delta\mathcal{Q}}) + 1} - i = m + 1 - \frac{2m+1}{\exp(\frac{\epsilon}{\Delta\mathcal{Q}}) + 1} - \Delta\mathcal{Q}$$

Let $\bar{B}$ be the random variable of shifted $Binomial(2m, \frac{1}{2})$ with zero mean and realization $\bar{b}$. According to Chernoff bound, $\forall t \in [0, \sqrt{2m}]$, there is $\Pr(\bar{B} \geq t\frac{\sqrt{2m}}{2}) \leq e^{-t^2/2}$.

Then if $1 \leq \Delta\mathcal{Q} + \frac{2m+1}{\exp(\frac{\epsilon}{\Delta\mathcal{Q}})+1} \leq m + 1$, there is:

$$\Pr(\bar{B} \geq m + 1 - \frac{2m+1}{\exp(\frac{\epsilon}{\Delta\mathcal{Q}}) + 1} - \Delta\mathcal{Q}) \leq \exp(-\frac{1}{m}(m - \Delta\mathcal{Q} + 1 - \frac{2m+1}{\exp(\frac{\epsilon}{\Delta\mathcal{Q}}) + 1})^2) = \delta$$

Similarly, given $\delta \in [0, 1]$, the corresponding ϵ is:

$$\epsilon = \Delta\mathcal{Q} \log(\frac{2m+1}{m - \Delta\mathcal{Q} + 1 - \sqrt{m \log \frac{1}{\delta}}} - 1)$$

B.3 Proof of Theorem 8

According to [4], for a mechanism $\mathscr{M}$ outputs o, with inputs d and $\hat{d}$, let a random variable $c(o;\mathscr{M},d,\hat{d}) = \log\frac{\Pr(\mathscr{M}(d)=o)}{\Pr(\mathscr{M}(\hat{d})=o)}$ denote the privacy loss at o, and

$$\alpha_{\mathscr{M}}(\lambda) = \max_{d,\hat{d}} \log \mathbb{E}_{o\sim\mathscr{M}(d)}\{\exp(\lambda c(o;\mathscr{M},d,\hat{d}))\}$$

There is:

$$\begin{aligned} & c(x_{1:T};\mathscr{M},z_{1:T},\hat{z}_{1:T}) = \log\frac{\mathscr{F}_{X_{1:T}|Z_{1:T}}(x_{1:T}|z_{1:T})}{\mathscr{F}_{X_{1:T}|Z_{1:T}}(x_{1:T}|z_{1:T})} \\ = & \sum_{t=2}^{T}\log\frac{\mathscr{F}_{X_t|Z_t,X_{1:t-1}}(x_t|z_t,x_{1:t-1})}{\mathscr{F}_{X_t|Z_t,X_{1:t-1}}(x_t|\hat{z}_t,x_{1:t-1})} + \log\frac{\mathscr{F}_{X_1|Z_1}(x_1|z_1)}{\mathscr{F}_{X_1|Z_1}(x_1|\hat{z}_1)} = \sum_{t=1}^{T} c(x_t;\mathscr{M}_t,z_t,\hat{z}_t) \end{aligned}$$

and for any pair of sequences $z_{1:T},\hat{z}_{1:T}$, the following holds

$$\begin{aligned} & \log\mathbb{E}_{X_{1:T}\sim\mathscr{M}(Z_{1:T})}\{\exp(\lambda c(x_{1:T};\mathscr{M},z_{1:T},\hat{z}_{1:T}))\} \\ = & \log\mathbb{E}_{X_{1:T}\sim\mathscr{M}(Z_{1:T})}\{\exp(\lambda\sum_{t=1}^{T} c(x_t;\mathscr{M}_t,z_t,\hat{z}_t)\} \\ \le & \sum_{t=1}^{T}\log\mathbb{E}_{X_t\sim\mathscr{M}(Z_t)}\{\exp(\lambda c(x_t;\mathscr{M}_t,z_t,\hat{z}_t)\} \end{aligned} \tag{B.3}$$

Therefore, $\alpha_{\mathscr{M}}(\lambda) \le \sum_{t=1}^{T}\alpha_{\mathscr{M}_t}(\lambda)$ also holds.

Consider $\alpha_{\mathscr{M}_t}(\lambda)$ first.

For $t \le 2-T_0$, $X_t = Z_t + N_t$ with $N_t \sim \mathcal{N}(0,\sigma_n^2)$

$$c(x_t;\mathscr{M}_t,z_t,\hat{z}_t) = \log\frac{\mathscr{F}_{X_t|Z_t,X_{1:t-1}}(x_t|z_t,x_{1:t-1})}{\mathscr{F}_{X_t|Z_t,X_{1:t-1}}(x_t|\hat{z}_t,x_{1:t-1})} = \log\frac{\mathscr{F}_{N_t}(n_t)}{\mathscr{F}_{N_t}(\hat{n}_t)} \le \frac{1}{2\sigma_n^2}\Delta(2n_t+\Delta)\ .$$

$$\begin{aligned} & \alpha_{\mathscr{M}_t}(\lambda) = \log\mathbb{E}_{N_t\sim\mathcal{N}(0,\sigma_n^2)}\{\exp(\lambda\frac{1}{2\sigma_n^2}\Delta(2n_t+\Delta))\} \\ = & \log\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma_n}\exp(-\frac{1}{2\sigma_n^2}(n_t-\lambda\Delta)^2)\cdot\exp(\frac{1}{2\sigma_n^2}(\lambda^2+\lambda)\Delta^2)dn_t = \frac{\lambda(\lambda+1)\Delta^2}{2\sigma_n^2}\ . \end{aligned}$$

For $t > 2$,

$$X_t = (1 - w_t)(\hat{\mu}_{t-1}(1 - r_t) + r_t X_{t-1}) + w_t Z_t + N_t$$

with $N_t \sim \mathcal{N}(0, \sigma_n^2)$.

$$c(x_t; \mathscr{M}_t, z_t, \hat{z}_t) = \log \frac{\mathscr{F}_{X_t|Z_t, X_{1:t-1}}(x_t|z_t, x_{1:t-1})}{\mathscr{F}_{X_t|Z_t, X_{1:t-1}}(x_t|\hat{z}_t, x_{1:t-1})} = \log \frac{\mathscr{F}_{N_t}(n_t)}{\mathscr{F}_{N_t}(\hat{n}_t)} \leq \frac{1}{2\sigma_n^2} w_t \Delta (2n_t + w_t \Delta) .$$

$$\begin{aligned} \alpha_{\mathscr{M}_t}(\lambda) &= \log \mathbb{E}\{\exp(\lambda \frac{1}{2\sigma_n^2} w_t \Delta (2n_t + w_t \Delta))\} \\ &= \log \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_n} \exp(-\frac{1}{2\sigma_n^2}(n_t - \lambda w_t \Delta)^2) \cdot \exp(\frac{1}{2\sigma_n^2}(\lambda^2 + \lambda) w_t^2 \Delta^2) dn_t \\ &= \frac{\lambda(\lambda + 1) w_t^2 \Delta^2}{2\sigma_n^2} . \end{aligned}$$

If let $w_t = 1$ for $t \leq 2$, there is: $\alpha_{\mathscr{M}}(\lambda) \leq \lambda(\lambda + 1) \frac{\Delta^2}{2\sigma_n^2} \sum_{t=1}^{T} w_t^2$. Use the tail bound [Theorem 2, [4]], for any $\epsilon_T \geq \frac{\Delta^2}{2\sigma_n^2} \sum_{t=1}^{T} w_t^2$, the algorithm is (ϵ_T, δ_T)-differentially private for

$$\delta_T = \min_{\lambda: \lambda \geq 0} h(\lambda) = \min_{\lambda: \lambda \geq 0} \exp(\lambda(\lambda + 1) \frac{\Delta^2}{2\sigma_n^2} \sum_{t=1}^{T} w_t^2 - \lambda \epsilon_T) \tag{B.4}$$

To find $\lambda^* = \underset{\lambda: \lambda \geq 0}{\operatorname{argmin}} h(\lambda)$, take derivative of $h(\lambda)$ and assign 0 gives the solution $\bar{\lambda} = \frac{\epsilon_T}{\frac{\Delta^2}{\sigma_n^2} \sum_{t=1}^{T} w_t^2} - \frac{1}{2} \geq 0$, and $h''(\bar{\lambda}) > 0$, implies $\lambda^* = \bar{\lambda}$. Plug into (B.4) gives:

$$\delta_T = \exp((\frac{\frac{\Delta^2}{\sigma_n^2} \sum_{t=1}^{T} w_t^2}{4} - \frac{\epsilon_T}{2})(\frac{\epsilon_T}{\frac{\Delta^2}{\sigma_n^2} \sum_{t=1}^{T} w_t^2} - \frac{1}{2})) \tag{B.5}$$

Similarly, for any $\delta_T \in [0, 1]$, the algorithm is (ϵ_T, δ_T)-differentially private for

$$\epsilon_T = \min_{\lambda: \lambda \geq 0} h_1(\lambda) = \min_{\lambda: \lambda \geq 0} (\lambda + 1) \frac{\Delta^2}{2\sigma_n^2} \sum_{t=1}^{T} w_t^2 + \frac{1}{\lambda} \log(\frac{1}{\delta_T}) \tag{B.6}$$

with $\lambda^* = \underset{\lambda:\lambda\geq 0}{\operatorname{argmin}}\, h_1(\lambda) = \sqrt{\frac{\log\frac{1}{\delta_T}}{\frac{\Delta^2}{2\sigma_n^2}\sum_{t=1}^T w_t^2}}$. Plug into (B.6) gives:

$$\epsilon_T = 2\sqrt{\frac{\Delta^2}{2\sigma_n^2}\sum_{t=1}^T w_t^2 \log(\frac{1}{\delta_T})} + \frac{\Delta^2}{2\sigma_n^2}\sum_{t=1}^T w_t^2 \tag{B.7}$$

B.4 Proof of Corollary 1

Let $\phi = \frac{\Delta^2\sum_{t=1}^T w_t^2}{\sigma_n^2}$, then according to Theorem 8,

$$\ln\delta_T = (\frac{\phi}{4} - \frac{\epsilon_T}{2})(\frac{\epsilon_T}{\phi} - \frac{1}{2})$$

reorganize gives:

$$\phi^2 + (8\ln\delta_T - 4\epsilon_T)\phi + 4\epsilon_T^2 = 0$$

$$\phi = 2\epsilon_T - 4\ln\delta_T \pm 4\sqrt{(\ln\delta_T)^2 - \epsilon_T\ln\delta_T}$$

Since $\epsilon_T \geq \frac{\phi}{2}$ must hold, only one case is possible.

$$\phi = \frac{\Delta^2\sum_{t=1}^T w_t^2}{\sigma_n^2} = 2\epsilon_T - 4\ln\delta_T - 4\sqrt{(\ln\delta_T)^2 - \epsilon_T\ln\delta_T}$$

Therefore,

$$\sigma_n^2 = \frac{\Delta^2\sum_{t=1}^T w_t^2}{2\epsilon_T + 4\ln\frac{1}{\delta_T} - 4\sqrt{(\ln\frac{1}{\delta_T})^2 + \epsilon_T\ln\frac{1}{\delta_T}}}$$

B.5 Proof of Theorem 9

The data of each individual here spans over T time steps, the total privacy loss is the accumulation of privacy loss from T time steps:

$$\frac{\mathscr{F}_{X_{1:T}|Z_{1:T}}(x_{1:T}|z_{1:T})}{\mathscr{F}_{X_{1:T}|Z_{1:T}}(x_{1:T}|\hat{z}_{1:T})} = \frac{\mathscr{F}_{X_1|Z_1}(x_1|z_1)}{\mathscr{F}_{X_1|Z_1}(x_1|\hat{z}_1)} \cdot \prod_{t=2}^{T} \frac{\mathscr{F}_{X_t|Z_t,X_{1:t-1}}(x_t|z_t,x_{1:t-1})}{\mathscr{F}_{X_t|Z_t,X_{1:t-1}}(x_t|\hat{z}_t,x_{1:t-1})}$$

If x_t is released under (ϵ_t,δ_t)-differential privacy at time t, then the total privacy loss can be calculated using advanced composition theorem below:

Theorem 28. *(Advanced composition theorem for differential privacy [76]) For any $\epsilon_k > 0$, $\delta_k \in [0,1]$ for $k \in \{1,2,\cdots,T\}$, and $\tilde{\delta} \in [0,1]$, the class of (ϵ_k,δ_k)-differentially private mechanisms satisfy $(\tilde{\epsilon}_{\tilde{\delta}}, 1-(1-\tilde{\delta})\prod_{k=1}^{T}(1-\delta_k))$-differential privacy under T-fold adaptive composition, for*

$$\begin{aligned}\tilde{\epsilon}_{\tilde{\delta}} \quad = \quad & \min\Big\{\sum_{k=1}^{T}\frac{(e^{\epsilon_k}-1)\epsilon_k}{e^{\epsilon_k}+1} + \sqrt{\sum_{k=1}^{T}2\epsilon_k^2\log(e+\frac{\sqrt{\sum_{k=1}^{T}\epsilon_k^2}}{\tilde{\delta}})}, \\ & \sum_{k=1}^{T}\epsilon_k, \sum_{k=1}^{T}\frac{(e^{\epsilon_k}-1)\epsilon_k}{e^{\epsilon_k}+1} + \sqrt{\sum_{k=1}^{T}2\epsilon_k^2\log(\frac{1}{\tilde{\delta}})}\Big\}\end{aligned}$$

First calculate the (ϵ_t,δ_t) at each stage by Lemma 4. Since $N_t + m \sim Binomial(2m, \frac{1}{2})$, for $t \le 2$, $X_t = Z_t + N_t$ so that the sensitivity $\Delta\mathcal{Q}_t = \Delta$; for $t > 2$, $X_t = [(1-w_t)(\hat{\mu}_{t-1}(1-r_t) + r_t X_{t-1}) + w_t Z_t + N_t]$ and the sensitivity is $\Delta\mathcal{Q}_t = w_t\Delta$. Let $w_t = 1$ for $t \le 2$. Then $\forall\epsilon_t > 0$,

$$\delta_t = \exp(-\frac{1}{m}(m - w_t\Delta + 1 - \frac{2m+1}{\exp(\frac{\epsilon_t}{w_t\Delta})+1})^2)$$

or $\forall\delta_t \in (0,1)$,

$$\epsilon_t = w_t\Delta\log(\frac{2m+1}{m - w_t\Delta + 1 - \sqrt{m\log\frac{1}{\delta_t}}} - 1)$$

Apply Theorem 28 directly, Theorem 9 is proved.

B.6 Proof of Theorem 10

$$\begin{aligned}\mathbb{E}_{X_{1:T}}(\|x_{1:T}-z_{1:T}\|^2) &= \mathbb{E}_{X_{1:T}}(\sum_{t=1}^{T}(x_t-z_t)^2)\\ &= \mathbb{E}_{X_{1:T-1}}\{\sum_{t=1}^{T-1}(x_t-z_t)^2+\underbrace{\mathbb{E}_{X_T|X_{1:T-1}}[(x_T-z_T)^2]}_{\textbf{term 1}}\}\end{aligned} \tag{B.8}$$

Replacing $x_T=(1-w_T)\hat{z}_T(x_{T-1})+w_Tz_T+n_T$ into **term 1** gives:

$$\begin{aligned}\textbf{term 1} &= \mathbb{E}_{X_T|X_{1:T-1}}[((1-w_t)(\hat{z}_T(x_{T-1})-z_T)+n_T)^2]\\ &= (1-w_T)^2(\hat{z}_T(x_{T-1})-z_T)^2+\sigma_n^2\end{aligned}$$

Plug into Eqn. (B.8):

$$\begin{aligned}\text{(B.8)} &= \mathbb{E}_{X_{1:T-1}}\{\sum_{t=1}^{T-1}(x_t-z_t)^2+(1-w_T)^2(\hat{z}_T(x_{T-1})-z_T)^2+\sigma_n^2\}\\ &= \mathbb{E}_{X_{1:T-2}}\{\sum_{t=1}^{T-2}(x_t-z_t)^2+\sigma_n^2+\textbf{term 2}\}\end{aligned}$$

with

$$\begin{aligned}\textbf{term 2} &= \mathbb{E}_{X_{T-1}|X_{1:T-2}}\{(x_{T-1}-z_{T-1})^2+(1-w_T)^2(\hat{z}_T(x_{T-1})-z_T)^2\}\\ &= (1-w_{T-1})^2(\hat{z}_{T-1}(x_{T-2})-z_{T-1})^2+\sigma_n^2+(1-w_T)^2\underbrace{\mathbb{E}_{X_{T-1}}\{(\hat{z}_T(x_{T-1})-z_T)^2\}}_{\textbf{term 3}}\end{aligned}$$

Since $\hat{z}_T(x_{T-1})$ is the LMMSE estimator of Z_T given X_{T-1}, **term 3** is just the corresponding MSE. For a Gaussian AR(1) process $Z_{1:T}$ with $Z_t\sim\mathcal{N}(\mu,\sigma_z^2)$ and $\text{Corr}(Z_tZ_{t-1})=\rho$. There is:

$$\textbf{term 3}=\sigma_z^2(1-\rho^2\frac{\sigma_z^2}{\sigma_z^2+\sigma_n^2})$$

Therefore,

$$
\begin{aligned}
(\text{B.8}) \quad &= \mathbb{E}_{X_{1:T-2}}\{\sum_{t=1}^{T-2}(x_t - z_t)^2 + (1-w_{T-1})^2(\hat{z}_{T-1}(x_{T-2}) - z_{T-1})^2\} \\
&\quad +(1-w_T)^2\sigma_z^2(1-\rho^2\frac{\sigma_z^2}{\sigma_z^2+\sigma_n^2}) + 2\sigma_n^2 \\
&= \cdots \\
&= \mathbb{E}_{X_1}\{(x_1 - z_1)^2 + (1-w_2)^2(\hat{z}_2(x_1) - z_2)^2\} \\
&\quad +\sigma_z^2(1-\rho^2\frac{\sigma_z^2}{\sigma_z^2+\sigma_n^2})\sum_{t=3}^{T}(1-w_t)^2 + (T-1)\sigma_n^2 \\
&= \sigma_z^2(1-\rho^2\frac{\sigma_z^2}{\sigma_z^2+\sigma_n^2})\sum_{t=2}^{T}(1-w_t)^2 + T\sigma_n^2
\end{aligned}
$$

Since $w_1 = 1$,

$$
\mathbb{E}_{X_{1:T}}(\|x_{1:T} - z_{1:T}\|^2) = \sigma_z^2(1-\rho^2\frac{\sigma_z^2}{\sigma_z^2+\sigma_n^2})\sum_{t=1}^{T}(1-w_t)^2 + T\sigma_n^2
$$

B.7 Proof of Theorem 11

Since both satisfy (ϵ_T, δ_T)-differential privacy, according to Corollary 1, $(\sigma_n^2)^A$, $(\sigma_n^2)^B$ should satisfy:

$$
\frac{T}{(\sigma_n^2)^B} = \frac{\sum_{t=1}^{T} w_t^2}{(\sigma_n^2)^A} = \frac{2\epsilon_T + 4\ln\frac{1}{\delta_T} - 4\sqrt{(\ln\frac{1}{\delta_T})^2 + \epsilon_T \ln\frac{1}{\delta_T}}}{\Delta^2}
$$

By Theorem 10,

$$
\begin{aligned}
\mathbb{E}_{X^A_{1:T}}(\|x_{1:T} - z_{1:T}\|^2) &= \sigma_z^2(1-\rho^2\frac{\sigma_z^2}{\sigma_z^2+(\sigma_n^2)^A})\sum_{t=1}^{T}(1-w_t)^2 + T(\sigma_n^2)^A \\
\mathbb{E}_{X^B_{1:T}}(\|x_{1:T} - z_{1:T}\|^2) &= T(\sigma_n^2)^B = T(\sigma_n^2)^A\frac{T}{\sum_{t=1}^{T} w_t^2}
\end{aligned}
$$

If $\mathbb{E}_{X^A_{1:T}}(\|x_{1:T} - z_{1:T}\|^2) < \mathbb{E}_{X^B_{1:T}}(\|x_{1:T} - z_{1:T}\|^2)$, then $\exists t$ s.t. $w_t \neq 1$ and

$$\sigma_z^2(1-\rho^2\frac{\sigma_z^2}{\sigma_z^2+(\sigma_n^2)^A})\sum_{t=1}^{T}(1-w_t)^2 < T(\sigma_n^2)^A(\frac{T}{\sum_{t=1}^{T}w_t^2}-1)$$

Reorganize it implies:

$$\frac{(\sigma_n^2)^A/\sigma_z^2}{1-\frac{\rho^2}{1+(\sigma_n^2)^A/\sigma_z^2}} > \frac{\sum_{t=1}^{T}(1-w_t)^2}{T(\frac{T}{\sum_{t=1}^{T}w_t^2}-1)} \tag{B.9}$$

Therefore, if $\exists\ \{w_t\}_{t=1}^T, w_t \in (0,1)$ and $(\sigma_n^2)^A$ satisfy both (B.9) and (B.9), then $x^A_{1:T}$ will be more accurate than $x^B_{1:T}$.

Consider the case when $w_t = w \in (0,1), \forall t$.

Then the right hand side of (B.9) is reduced to $h_1(w) = \frac{(1-w)^2}{\frac{1}{w^2}-1}$, since

$$\begin{aligned}\lim_{w\to 1} h_1(w) &= 0\\ \lim_{w\to 0} h_1(w) &= 0\\ h_1'(w) &= \frac{-2w(w^2+w-1)}{(1+w)^2}\end{aligned}$$

$\exists$ only one $\bar{w}$ over $(0,1)$ such that $\bar{w}^2+\bar{w}-1=0$. Therefore, $h_1(w)$ is strictly increasing from 0 to $h_1(\bar{w})>0$ over $(0,\bar{w})$ and strictly decreasing over from $h_1(\bar{w})>0$ to 0 over $(\bar{w},1)$.

Let $\xi = (\sigma_n^2)^A/\sigma_z^2 \geq 0$, then the left hand side of (B.9) can be re-written as $h_2(\xi) = \frac{\xi}{1-\frac{\rho^2}{1+\xi}}$, we have:

$$h_2'(\xi) = \frac{\xi^2+2\xi(1-\rho^2)+(1-\rho^2)}{(1+\xi-\rho^2)^2}$$

Since $h_2(0)=0$ and $h_2'(\xi)>0$ over $\xi\in[0,\infty)$, $h_2(\xi)$ is strictly increasing from 0 to $+\infty$ over $\xi\in[0,\infty)$. For all pairs of $(w,(\sigma_n^2)^A)$ satisfying (B.9), w and $(\sigma_n^2)^A$ is bijective and we can write $\xi = h_3(w)$ for some strictly increasing function h_3.

Since both h_2, h_3 are strictly increasing functions, $h_2(h_3(w))$ is strictly increasing from 0 over $w\in(0,1)$. Therefore, $\exists w\in(0,1)$, such that $h_2(h_3(w)) > h_1(w)$ and $x^A_{1:T}$ released by our method is more accurate than $x^B_{1:T}$.

Moreover, if $w > \frac{1-(\sigma_n^2)^B/\sigma_z^2}{1+(\sigma_n^2)^B/\sigma_z^2}$, then re-organize it implies

$$w^2 \frac{(\sigma_n^2)^B}{\sigma_z^2} > h_1(w).$$

Since $h_2(h_3(w)) = \frac{w^2 \frac{(\sigma_n^2)^B}{\sigma_z^2}}{1-\frac{\rho^2}{1+w^2 \frac{(\sigma_n^2)^B}{\sigma_z^2}}} > w^2 \frac{(\sigma_n^2)^B}{\sigma_z^2}$, it further implies $h_2(h_3(w)) > h_1(w)$.

Therefore, if

$$w > \frac{1-(\sigma_n^2)^B/\sigma_z^2}{1+(\sigma_n^2)^B/\sigma_z^2},$$

then $x_{1:T}^A$ will be more accurate than $x_{1:T}^B$.

APPENDIX C

Long-Term Impact of Fairness Interventions on Group Representation

C.1 Proof of Theorem 12

Theorem 12 is proved based on the following Lemma.

Lemma 15. *Let a, b, z_a, z_b be real constants, where $a, b \in \mathbb{R}_+$ and $z_a, z_b \in [0,1]$. If $b \geq a > 1$, $z_b - z_a > \frac{1}{a} - \frac{1}{b}$ and $b < \frac{1}{1-z_b}$ are satisfied, then the following holds:*

$$\frac{1 + z_a + a z_a^2}{1 + z_b + b z_b^2} \leq \frac{1 + a z_a}{1 + b z_b} \tag{C.1}$$

Proof. Re-organizing (C.1) gives the following:

$$\begin{aligned} (1 + z_a + a z_a^2)(1 + b z_b) &\leq (1 + z_b + b z_b^2)(1 + a z_a) \\ b z_b + b z_a z_b + z_a + a b z_a^2 z_b + a z_a^2 &\leq a z_a + z_b + a z_a z_b + b z_b^2 + a b z_b^2 z_a \end{aligned}$$

Proving (C.1) is equivalent to showing the following:

$$0 \leq (a-1)\frac{1}{z_b} + (1-b)\frac{1}{z_a} + b\frac{z_b}{z_a} - a\frac{z_a}{z_b} + \underbrace{a - b + ab(z_b - z_a)}_{\textbf{term 1}}$$

Since $z_b - z_a > \frac{1}{a} - \frac{1}{b}$, **term 1** $> a - b + b - a = 0$ holds. Therefore, proving (C.1) is equivalent to

showing:

$$az_a^2 + (1-a)z_a \leq bz_b^2 + (1-b)z_b \tag{C.2}$$

Since $b < \frac{1}{1-z_b}$ holds, implying $z_b > 1 - \frac{1}{b}$.

Define a function $g(z) = cz^2 + (1-c)z$, $z \in [0,1]$ under any constant $c > 1$. The following holds:

$$g(1-\frac{1}{c}) = 0; \qquad g(1) = 1$$
$$g'(z) = 2cz + 1 - c; \quad g'(1-\tfrac{1}{c}) = c - 1; \quad g''(z) = 2c$$

Since $g''(z)$ is a positive constant over $z \in [0,1]$, $g'(z)$ is strictly increasing and $g'(z) > 0$ when $z \in (1-\frac{1}{c}, 1]$, thus $g(z)$ is increasing over $z \in (1-\frac{1}{c}, 1]$ from 0 to 1.

Now consider two functions $g_a(z) = az^2 + (1-a)z$ and $g_b(z) = bz^2 + (1-b)z$ with $z \in [0,1]$. From the above analysis, $g_a(z)$ is increasing over $(1-\frac{1}{a}, 1]$ from 0 to 1 and $g_b(z)$ is increasing over $(1-\frac{1}{b}, 1]$ from 0 to 1. Moreover, $1-\frac{1}{b} \geq 1-\frac{1}{a}$ and $g_b''(z) = 2b \geq 2a = g_a''(z)$, i.e., the speed that $g_b(z)$ increases over $(1-\frac{1}{b}, 1]$ is NOT slower than the speed that $g_a(z)$ increases over $(1-\frac{1}{a}, 1]$. Since $z_b - z_a > \frac{1}{a} - \frac{1}{b} = (1-\frac{1}{b}) - (1-\frac{1}{a})$ and $z_b > 1 - \frac{1}{b}$, $g_a(z_a) \leq g_b(z_b)$ must hold.

Therefore, (C.2) is satisfied. Inequality (C.1) is proved.

□

To simplify the notation, denote $\lambda_{k,t} := \lambda_{k,t}(\theta_k(t))$. We will only present the case when $\diamond :=$ " $<$ ", cases when $\diamond :=$ " $>$ " and $\diamond :=$ " $=$ " can be derived similarly and are omitted.

To prove Theorem 12, we prove the following statement using induction: If $\lambda_{a,1} < \lambda_{b,1}$, then $\forall t$, $\frac{n_a(t+1)}{n_b(t+1)} < \frac{n_a(t)}{n_b(t)}$ and $\lambda_{a,t+1} < \lambda_{a,t} < \lambda_{b,t} < \lambda_{b,t+1}$ hold under monotonicity condition. Moreover, $N_b(t) < \frac{\beta_b}{1-\lambda_{b,t}}, \forall t$.

Base Case:

Since $\frac{N_a(1)}{N_b(1)} = \frac{\beta_a}{\beta_b}$. If $\lambda_{a,1} < \lambda_{b,1}$, then $\frac{n_a(2)}{n_b(2)} = \frac{N_a(1)\lambda_{a,1}+\beta_a}{N_b(1)\lambda_{b,1}+\beta_b} < \frac{N_a(1)}{N_b(1)} = \frac{n_a(1)}{n_b(1)}$. Under monotonicity condition, it results in $\lambda_{a,2} < \lambda_{a,1} < \lambda_{b,1} < \lambda_{b,2}$. Moreover, since $N_b(2) = N_b(1)\lambda_{b,1} + \beta_b > N_b(1)$, implying $N_b(1) < \frac{\beta_b}{1-\lambda_{b,1}}$.

Induction Step:

Suppose $\frac{n_a(t+1)}{n_b(t+1)} < \frac{n_a(t)}{n_b(t)} \leq \frac{\beta_a}{\beta_b}$, $\lambda_{a,t+1} < \lambda_{a,t} < \lambda_{b,t} < \lambda_{b,t+1}$ and $N_b(t) < \frac{\beta_b}{1-\lambda_{b,t}}$ hold at time $t \geq 1$. Show that for time step $t+1$, $\frac{n_a(t+2)}{n_b(t+2)} < \frac{n_a(t+1)}{n_b(t+1)} \leq \frac{\beta_a}{\beta_b}$, $\lambda_{a,t+2} < \lambda_{a,t+1} < \lambda_{b,t+1} < \lambda_{b,t+2}$ and $N_b(t+1) < \frac{\beta_b}{1-\lambda_{b,t+1}}$ also hold.

Denote $N_a(t) = c_a\beta_a$ and $N_b(t) = c_b\beta_b$. Since $N_k(t) = N_k(t-1)\lambda_{k,t-1} + \beta_k > \beta_k, \forall t$, it holds that $c_a, c_b > 1$.

By hypothesis, $\frac{n_a(t)}{n_b(t)} \leq \frac{\beta_a}{\beta_b}$ implies that $c_b \geq c_a > 1$, and $N_b(t) < \frac{\beta_b}{1-\lambda_{b,t}}$ implies that $c_b < \frac{1}{1-\lambda_{b,t}}$. Since $\frac{N_a(t+1)}{N_b(t+1)} = \frac{N_a(t)\lambda_{a,t}+\beta_a}{N_b(t)\lambda_{b,t}+\beta_b} = \frac{\beta_a}{\beta_b}\frac{c_a\lambda_{a,t}+1}{c_b\lambda_{b,t}+1} < \frac{N_a(t)}{N_b(t)} = \frac{\beta_a}{\beta_b}\frac{c_a}{c_b}$, re-organizing it gives $\lambda_{b,t} - \lambda_{a,t} > \frac{1}{c_a} - \frac{1}{c_b}$.

By Lemma 15, the following holds:

$$\frac{N_a(t)\lambda_{a,t}^2 + \beta_a(1+\lambda_{a,t})}{N_b(t)\lambda_{b,t}^2 + \beta_b(1+\lambda_{b,t})} = \frac{\beta_a}{\beta_b}\frac{1+\lambda_{a,t}+c_a\lambda_{a,t}^2}{1+\lambda_{b,t}+c_b\lambda_{b,t}^2} \leq \frac{\beta_a}{\beta_b}\frac{1+c_a\lambda_{a,t}}{1+c_b\lambda_{b,t}} = \frac{N_a(t+1)}{N_b(t+1)} = \frac{n_a(t+1)}{n_b(t+1)}$$

Since we suppose $\lambda_{a,t+1} < \lambda_{a,t} < \lambda_{b,t} < \lambda_{b,t+1}$, we have:

$$\frac{N_a(t)\lambda_{a,t}^2 + \beta_a(1+\lambda_{a,t})}{N_b(t)\lambda_{b,t}^2 + \beta_b(1+\lambda_{b,t})} > \frac{(N_a(t)\lambda_{a,t}+\beta_a)\lambda_{a,t+1}+\beta_a}{(N_b(t)\lambda_{b,t}+\beta_b)\lambda_{b,t+1}+\beta_b} = \frac{n_a(t+2)}{n_b(t+2)}$$

It implies that $\frac{n_a(t+2)}{n_b(t+2)} < \frac{n_a(t+1)}{n_b(t+1)}$.

By motonoticity condition, it results in $\lambda_{a,t+2} < \lambda_{a,t+1} < \lambda_{b,t+1} < \lambda_{b,t+2}$.

Moreover, $N_b(t+1) = N_b(t)\lambda_{b,t} + \beta_b < \frac{\beta_b\lambda_{b,t}}{1-\lambda_{b,t}} + \beta_b = \frac{\beta_b}{1-\lambda_{b,t}} < \frac{\beta_b}{1-\lambda_{b,t+1}}$.

The statement holds for time $t+1$. This completes the proof.

C.2 Proof of Theorem 13

Without loss of generality, let $\frac{\widehat{n}_a}{\widehat{n}_b} < \frac{\widetilde{n}_a}{\widetilde{n}_b}$. Since $\lambda_k(\theta_k) = h_k(O_k(\theta_k))$ with $h_k(\cdot)$ being a decreasing function, showing that $\widetilde{\mathbf{O}}$ and $\widehat{\mathbf{O}}$ satisfy Monotonicity condition is equivalent to showing that $O_a(\widehat{\theta}_a) > O_a(\widetilde{\theta}_a)$, $O_b(\widehat{\theta}_b) < O_b(\widetilde{\theta}_b)$. Under the condition that $O_k(\widehat{\theta}_k) \neq O_k(\widetilde{\theta}_k)$ for any possible $\widehat{n}_a \neq \widetilde{n}_a$, prove by contradiction: suppose $O_a(\widehat{\theta}_a) < O_a(\widetilde{\theta}_a)$ holds, then $O_b(\widehat{\theta}_b) > O_b(\widetilde{\theta}_b)$ must also hold otherwise $(\widehat{\theta}_a, \widehat{\theta}_b)$ will be the solution to $\widetilde{\mathbf{O}}$.

Because $(\widehat{\theta}_a, \widehat{\theta}_b)$ is the optimal solution to $\widehat{\mathbf{O}}$ and $(\widetilde{\theta}_a, \widetilde{\theta}_b)$ is the optimal solution to $\widetilde{\mathbf{O}}$, and $O_b(\widehat{\theta}_b) > O_b(\widetilde{\theta}_b)$, the following holds:

$$\widehat{n}_a O_a(\widehat{\theta}_a) + \widehat{n}_b O_b(\widehat{\theta}_b) \leq \widehat{n}_a O_a(\widetilde{\theta}_a) + \widehat{n}_b O_b(\widetilde{\theta}_b) \rightarrow \frac{O_a(\widehat{\theta}_a) - O_a(\widetilde{\theta}_a)}{O_b(\widetilde{\theta}_b) - O_b(\widehat{\theta}_b)} \geq \frac{\widehat{n}_b}{\widehat{n}_a}$$

$$\widetilde{n}_a O_a(\widetilde{\theta}_a) + \widetilde{n}_b O_b(\widetilde{\theta}_b) \leq \widetilde{n}_a O_a(\widehat{\theta}_a) + \widetilde{n}_b O_b(\widehat{\theta}_b) \rightarrow \frac{O_a(\widehat{\theta}_a) - O_a(\widetilde{\theta}_a)}{O_b(\widetilde{\theta}_b) - O_b(\widehat{\theta}_b)} \leq \frac{\widetilde{n}_b}{\widetilde{n}_a}$$

It implies that $\frac{\widehat{n}_a}{\widehat{n}_b} \geq \frac{\widetilde{n}_a}{\widetilde{n}_b}$, which is a contradiction.

C.3 Proof of Lemma 5

Starting from Appendix C.3 until Appendix C.6, we simplify the notations by removing t from subscript, i.e., $L_{s,t}(\theta_s) := L_s(\theta_s)$, $\alpha_{s,t} := \alpha_s$, $f_{s,t}(x) := f_s(x)$, $f^y_{s,t}(x) := f^y_s(x)$, $\underline{s}^y_t := \underline{s}^y$, $\overline{s}^y_t := \overline{s}^y$, $\eta^{\mathcal{C}}_t := \eta^{\mathcal{C}}$, $\Gamma_{\mathcal{C},t} := \Gamma_{\mathcal{C}}$, $\delta_{s,t} := \delta_s$, $\mathcal{T}_{s,t} := \mathcal{T}_s$.

The loss for group s can be written as

$$\begin{aligned} L_s(\theta_s) &= \int_{-\infty}^{\theta_s} \alpha_s f^1_s(x)dx + \int_{\theta_s}^{\infty} (1-\alpha_s) f^0_s(x)dx \\ &= \begin{cases} \int_{\theta_s}^{\overline{s}^0} (1-\alpha_s) f^0_s(x)dx, \text{ if } \theta_s \in [\underline{s}^0, \underline{s}^1] \\ \int_{\theta_s}^{\overline{s}^0} (1-\alpha_s) f^0_s(x)dx + \int_{\underline{s}^1}^{\theta_s} \alpha_s f^1_s(x)dx, \text{ if } \theta_s \in [\underline{s}^1, \overline{s}^0] \\ \int_{\underline{s}^1}^{\theta_s} \alpha_s f^1_s(x)dx, \text{ if } \theta_s \in [\overline{s}^0, \overline{s}^1] \end{cases} \end{aligned}$$

which is decreasing in θ_s over $[\underline{s}^0, \underline{s}^1]$ and increasing over $[\overline{s}^0, \overline{s}^1]$, the optimal solution $\theta^*_s \in [\underline{s}^1, \overline{s}^0]$. Taking derivative of $L_s(\theta_s)$ w.r.t. θ_s gives $\frac{dL_s(\theta_s)}{d\theta_s} = \alpha_s f^1_s(\theta_s) - (1-\alpha_s) f^0_s(\theta_s)$, which is strictly increasing over $[\underline{s}^1, \overline{s}^0]$ under Assumption 7.

The optimal solution $\theta^*_s = \arg\min_{\theta_s} L_s(\theta_s) \in \{\underline{s}^1, \delta_s, \overline{s}^0\}$ can be thus found easily. Moreover, $L_s(\theta_s)$ is decreasing in θ_s over $[\underline{s}^0, \theta^*_s]$ and increasing over $[\theta^*_s, \overline{s}^1]$.

C.4 Proof of Lemma 6

Some notations are simplified by removing subscript t as mentioned in Appendix C.3.

We proof this Lemma by contradiction.

Let $\mathcal{V} = \{(\theta_a, \theta_b) | \theta_a \in [\eta^{\mathcal{C}}(\delta_b), \delta_a], \theta_b \in [\delta_b, (\eta^{\mathcal{C}})^{-1}(\delta_a)], \Gamma_{\mathcal{C}}(\theta_a, \theta_b) = 0\}$.

Note that for `Simple`, `EqOpt`, `DP` fairness, for any (θ_a, θ_b) and (θ'_a, θ'_b) that satisfy constraints $\Gamma_{\mathcal{C}}(\theta_a, \theta_b) = 0$ and $\Gamma_{\mathcal{C}}(\theta'_a, \theta'_b) = 0$, $\theta_a \geq \theta'_a$ if and only if $\theta_b \geq \theta'_b$. Suppose that $(\check{\theta}_a, \check{\theta}_b)$ satisfies $\Gamma_{\mathcal{C}}(\check{\theta}_a, \check{\theta}_b) = 0$ and $(\check{\theta}_a, \check{\theta}_b) = \arg\min_{\theta_a, \theta_b} n_a L_a(\theta_a) + n_b L_b(\theta_b) \notin \mathcal{V}$, then one of the following must hold: (1) $\check{\theta}_a < \eta^{\mathcal{C}}(\delta_b)$, $\check{\theta}_b < \delta_b$; (2) $\check{\theta}_a > \delta_a$, $\check{\theta}_b > (\eta^{\mathcal{C}})^{-1}(\delta_a)$. Consider two cases separately.

(1) $\check{\theta}_a < \eta^{\mathcal{C}}(\delta_b)$, $\check{\theta}_b < \delta_b$

Since $L_b(\check{\theta}_b) > L_b(\delta_b)$, $\forall n_a, n_b$, to satisfy $n_a L_a(\check{\theta}_a) + n_b L_b(\check{\theta}_b) < n_a L_a(\eta^{\mathcal{C}}(\delta_b)) + n_b L_b(\delta_b)$, $L_a(\check{\theta}_a) < L_a(\eta^{\mathcal{C}}(\delta_b))$ must hold. However, by Lemma 5, $L_a(\theta_a)$ is strictly decreasing on $[\underline{a}^0, \delta_a]$ and strictly increasing on $[\delta_a, \overline{a}^1]$. Since $\check{\theta}_a < \eta^{\mathcal{C}}(\delta_b) < \delta_a$, this implies $L_a(\check{\theta}_a) > L_a(\eta^{\mathcal{C}}(\delta_b))$. Therefore, $(\check{\theta}_a, \check{\theta}_b)$ cannot be the optimal pair.

(2) $\check{\theta}_a > \delta_a$, $\check{\theta}_b > (\eta^{\mathcal{C}})^{-1}(\delta_a)$

Since $L_a(\check{\theta}_a) > L_a(\delta_a)$, $\forall n_a, n_b$, to satisfy $n_a L_a(\check{\theta}_a) + n_b L_b(\check{\theta}_b) < n_a L_a(\delta_a) + n_b L_b((\eta^{\mathcal{C}})^{-1}(\delta_a))$, $L_b(\check{\theta}_b) < L_b((\eta^{\mathcal{C}})^{-1}(\delta_a))$ must hold. However, by Lemma 5, $L_b(\theta_b)$ is strictly decreasing on $[\underline{b}^0, \delta_b]$ and strictly increasing on $[\delta_b, \overline{b}^1]$. Since $\check{\theta}_b > (\eta^{\mathcal{C}})^{-1}(\delta_a) > \delta_b$, this implies $L_b(\check{\theta}_b) > L_b((\eta^{\mathcal{C}})^{-1}(\delta_a))$. Therefore, $(\check{\theta}_a, \check{\theta}_b)$ cannot be the optimal pair.

C.5 Proof of Theorem 14

Some notations are simplified by removing subscript t as mentioned in Appendix C.3.

Proof of Theorem 14 is based on the following Lemma.

Lemma 16. *Consider the one-shot problem* (4.1) *at some time step t, with group proportions given by $n_a(t), n_b(t)$. Under Assumption 7 the one-shot decision $(\theta_a(t), \theta_b(t))$ for this time step is unique and satisfies the following:*

(1) Under `EqOpt` *fairness:*

- *If* $\theta_a(t) \in [\underline{a}^0, \underline{a}^1]$, $\theta_b(t) \in [\underline{b}^1, \overline{b}^0]$, *then* $\frac{n_a(t)}{n_b(t)} = (\frac{\alpha_b}{1-\alpha_b} \frac{f_b^1(\theta_b(t))}{f_b^0(\theta_b(t))} - 1) \frac{1-\alpha_b}{1-\alpha_a}$.
- *If* $\theta_a(t) \in [\underline{a}^1, \overline{a}^0]$, $\theta_b(t) \in [\underline{b}^1, \overline{b}^0]$, *then* $\frac{n_a(t)}{n_b(t)} = \frac{\frac{\alpha_b}{1-\alpha_b} \frac{f_b^1(\theta_b(t))}{f_b^0(\theta_b(t))} - 1}{1 - \frac{\alpha_a}{1-\alpha_a} \frac{f_a^1(\theta_a(t))}{f_a^0(\theta_a(t))}} \frac{1-\alpha_b}{1-\alpha_a}$.

(2) Under `DP` *fairness:*

- *If* $\theta_a(t) \in [\underline{a}^0, \underline{a}^1]$, $\theta_b(t) \in [\underline{b}^1, \overline{b}^0]$, *then* $\frac{n_a(t)}{n_b(t)} = 1 - \frac{2}{\frac{\alpha_b}{1-\alpha_b} \frac{f_b^1(\theta_b(t))}{f_b^0(\theta_b(t))} + 1}$.
- *If* $\theta_a(t) \in [\underline{a}^1, \overline{a}^0]$, $\theta_b(t) \in [\underline{b}^1, \overline{b}^0]$, *then* $\frac{n_a(t)}{n_b(t)} = (1 - \frac{2}{\frac{\alpha_b}{1-\alpha_b} \frac{f_b^1(\theta_b(t))}{f_b^0(\theta_b(t))} + 1})(\frac{2}{1 - \frac{\alpha_a}{1-\alpha_a} \frac{f_a^1(\theta_a(t))}{f_a^0(\theta_a(t))}} - 1)$.
- *If* $\theta_a(t) \in [\underline{a}^1, \overline{a}^0]$, $\theta_b(t) \in [\overline{b}^0, \overline{b}^1]$, *then* $\frac{n_a(t)}{n_b(t)} = \frac{2}{1 - \frac{\alpha_a}{1-\alpha_a} \frac{f_a^1(\theta_a(t))}{f_a^0(\theta_a(t))}} - 1$.

(3) Under `Simple` *fairness:*

- *If we further assume* $\delta_a, \delta_b \in \mathcal{T}_a \cap \mathcal{T}_b$, *then* $\theta_a(t) = \theta_b(t) \in [\underline{a}^1, \overline{b}^0]$ *and* $\frac{n_a(t)}{n_b(t)} = \frac{\alpha_b f_b^1(\theta_b(t)) - (1-\alpha_b) f_b^0(\theta_b(t))}{(1-\alpha_a) f_a^0(\theta_a(t)) - \alpha_a f_a^1(\theta_a(t))}$.

Proof. We focus on the case when $\alpha_a f_a^1(\underline{a}^1) < (1-\alpha_a) f_a^0(\underline{a}^1)$ & $\alpha_a f_a^1(\overline{a}^0) > (1-\alpha_a) f_a^0(\overline{a}^0)$ and $\alpha_b f_b^1(\underline{b}^1) < (1-\alpha_b) f_b^0(\underline{b}^1)$ & $\alpha_b f_b^1(\overline{b}^0) > (1-\alpha_b) f_b^0(\overline{b}^0)$. That is, $\theta_s^* = \text{argmin}_\theta L_s(\theta) = \delta_s$ holds for $s \in \{a, b\}$.

Constraint $\Gamma_{\mathcal{C}}(\theta_a, \theta_b) = 0$ can be rewritten as $\theta_a = \eta^{\mathcal{C}}(\theta_b)$ for some strictly increasing function $\eta^{\mathcal{C}}$. The following holds:

$$\frac{d\eta^{\mathcal{C}}(\theta_b)}{d\theta_b} = -\frac{\frac{\partial \Gamma_{\mathcal{C}}(\theta_a,\theta_b)}{\partial \theta_b}}{\frac{\partial \Gamma_{\mathcal{C}}(\theta_a,\theta_b)}{\partial \theta_a}}\Big|_{\theta_a=\eta^{\mathcal{C}}(\theta_b)} = \begin{cases} \frac{f_b^0(\theta_b)}{f_a^0(\eta^{\mathcal{C}}(\theta_b))}, \mathcal{C} := \texttt{EqOpt} \\ \frac{(1-\alpha_b) f_b^0(\theta_b) + \alpha_b f_b^1(\theta_b)}{(1-\alpha_a) f_a^0(\eta^{\mathcal{C}}(\theta_b)) + \alpha_a f_a^1(\eta^{\mathcal{C}}(\theta_b))}, \mathcal{C} := \texttt{DP} \\ 1, \mathcal{C} := \texttt{Simple} \end{cases}$$

The one-shot problem can be expressed with only one variable, either θ_a or θ_b. Here we express it in terms of θ_b. At each round, decision maker finds $\theta_b(t) = \arg\min_{\theta_b} n_a(t) L_a(\eta^{\mathcal{C}}(\theta_b)) + n_b(t) L_b(\theta_b)$ and $\theta_a(t) = \eta^{\mathcal{C}}(\theta_b(t))$. Since $\eta^{\mathcal{C}}(\delta_b) < \delta_a$ ($(\eta^{\mathcal{C}})^{-1}(\delta_a) > \delta_b$), when $\mathcal{C} := \texttt{DP}$, solution $(\theta_a(t), \theta_b(t))$ can be in one of the following three forms: (1) $\theta_a(t) \in [\underline{a}^0, \underline{a}^1]$, $\theta_b(t) \in [\underline{b}^1, \overline{b}^0]$; (2) $\theta_a(t) \in [\underline{a}^1, \overline{a}^0]$, $\theta_b(t) \in [\underline{b}^1, \overline{b}^0]$; (3) $\theta_a(t) \in [\underline{a}^1, \overline{a}^0]$, $\theta_b(t) \in [\overline{b}^0, \overline{b}^1]$. When $\mathcal{C} := \texttt{EqOpt}$, solution $(\theta_a(t), \theta_b(t))$ can be either (1) or (2) listed above. In the following analysis, we simplify the notation $\eta^{\mathcal{C}}$ as η when fairness criterion $\mathcal{C}$ is explicitly stated. For `EqOpt` and `DP` criteria, we consider each case separately.

Case 1: $\theta_a(t) \in [\underline{a}^0, \underline{a}^1]$, $\theta_b(t) \in [\underline{b}^1, \overline{b}^0]$

Let $\theta_b^{\max} = \min\{\overline{b}^0, (\eta^{\mathcal{C}})^{-1}(\underline{a}^1)\}$ be the maximum value θ_b can take.

$L^t(\theta_b) = n_b(t) \int_{\underline{b}^1}^{\theta_b} \alpha_b f_b^1(x) - (1-\alpha_b) f_b^0(x) dx - n_a(t) \int_{\underline{a}^0}^{\eta^{\mathcal{C}}(\theta_b)} (1-\alpha_a) f_a^0(x) dx + n_a(t)(1-\alpha_a) + n_b(t) \int_{\underline{b}^1}^{\overline{b}^0} (1-\alpha_b) f_b^0(x) dx$

Taking derivative w.r.t. θ_b gives

$$\frac{dL^t(\theta_b)}{d\theta_b} = n_b(t)(\alpha_b f_b^1(\theta_b) - (1-\alpha_b) f_b^0(\theta_b)) - n_a(t)(1-\alpha_a) f_a^0(\eta^{\mathcal{C}}(\theta_b)) \frac{d\eta^{\mathcal{C}}(\theta_b)}{d\theta_b}.$$

1. $\mathcal{C} := \texttt{EqOpt}$

$\frac{dL^t(\theta_b)}{d\theta_b} = n_b(t)(\alpha_b f_b^1(\theta_b) - (1-\alpha_b) f_b^0(\theta_b)) - n_a(t)(1-\alpha_a) f_b^0(\theta_b)$, since $\alpha_b f_b^1(\theta_b) - (1-\alpha_b) f_b^0(\theta_b)$ is increasing from negative to positive and $f_b^0(\theta_b)$ is decreasing over $[\underline{b}^1, \overline{b}^0]$, implying $\frac{dL^t(\theta_b)}{d\theta_b}$ is increasing over $[\underline{b}^1, \overline{b}^0]$. Based on the value of $\frac{n_a(t)}{n_b(t)}$,

• If $\frac{dL^t(\theta_b)}{d\theta_b}|_{\theta_b=\theta_b^{\max}} \geq 0$, then one-shot decision $\theta_b(t)$ satisfies $\frac{n_a(t)}{n_b(t)} = (\frac{\alpha_b}{1-\alpha_b} \frac{f_b^1(\theta_b(t))}{f_b^0(\theta_b(t))} - 1) \frac{1-\alpha_b}{1-\alpha_a}$ and is unique.

• If $\frac{dL^t(\theta_b)}{d\theta_b} < 0, \forall \theta_b \in [\underline{b}^1, \theta_b^{\max}]$, then $\theta_b(t) > \theta_b^{\max}$ and $(\theta_a(t), \theta_b(t))$ does not satisfy Case 1.

2. $\mathcal{C} :=$ DP

$\frac{dL^t(\theta_b)}{d\theta_b} = n_b(t)(\alpha_b f_b^1(\theta_b) - (1-\alpha_b) f_b^0(\theta_b)) - n_a(t)\frac{\alpha_b f_b^1(\theta_b) + (1-\alpha_b) f_b^0(\theta_b)}{1 + \frac{\alpha_a f_a^1(\eta(\theta_b))}{(1-\alpha_a) f_a^0(\eta(\theta_b))}} = (n_b(t) - n_a(t))\alpha_b f_b^1(\theta_b) - (n_b(t) + n_a(t))(1-\alpha_b) f_b^0(\theta_b)$, where the last equality holds since $f_a^1(\eta(\theta_b)) = 0$ over $[\underline{a}^0, \underline{a}^1]$. Since $\frac{dL^t(\theta_b)}{d\theta_b}|_{\theta_b = \underline{b}^1} < 0$, based on the value of $\frac{n_a(t)}{n_b(t)}$,

• If $\exists \theta_b'$ such that $\frac{dL^t(\theta_b)}{d\theta_b}|_{\theta_b = \theta_b'} \geq 0$, then one-shot decision $\theta_b(t)$ satisfies $\frac{n_a(t)}{n_b(t)} = 1 - \frac{2}{\frac{\alpha_b}{1-\alpha_b}\frac{f_b^1(\theta_b(t))}{f_b^0(\theta_b(t))} + 1}$ and is unique.

• If $\frac{dL^t(\theta_b)}{d\theta_b} < 0, \forall \theta_b \in [\underline{b}^1, \theta_b^{\max}]$, then $\theta_b(t) > \theta_b^{\max}$ and $(\theta_a(t), \theta_b(t))$ does not satisfy Case 1.

Case 2: $\theta_a(t) \in [\underline{a}^1, \overline{a}^0]$, $\theta_b(t) \in [\underline{b}^1, \overline{b}^0]$

Let $\theta_b^{\max} = \min\{\overline{b}^0, (\eta^{\mathcal{C}})^{-1}(\overline{a}^0)\}$ and $\theta_b^{\min} = \max\{\underline{b}^1, (\eta^{\mathcal{C}})^{-1}(\underline{a}^1)\}$ be the maximum and minimum value that θ_b can take respectively. $L^t(\theta_b) = n_b(t)\int_{\underline{b}^1}^{\theta_b} \alpha_b f_b^1(x) - (1-\alpha_b) f_b^0(x) dx + n_a(t)\int_{\underline{a}^1}^{\eta^{\mathcal{C}}(\theta_b)} \alpha_a f_a^1(x) - (1-\alpha_a) f_a^0(x) dx + n_b(t)\int_{\underline{b}^1}^{\overline{b}^0} (1-\alpha_b) f_b^0(x) dx + n_a(t)\int_{\underline{a}^1}^{\overline{a}^0} (1-\alpha_a) f_a^0(x) dx$

Taking derivative w.r.t. θ_b gives

$$\frac{dL^t(\theta_b)}{d\theta_b} = n_b(t)(\alpha_b f_b^1(\theta_b) - (1-\alpha_b) f_b^0(\theta_b)) + n_a(t)(\alpha_a f_a^1(\eta^{\mathcal{C}}(\theta_b)) - (1-\alpha_a) f_a^0(\eta^{\mathcal{C}}(\theta_b)))\frac{d\eta^{\mathcal{C}}(\theta_b)}{d\theta_b}.$$

1. $\mathcal{C} :=$ EqOpt

$\frac{dL^t(\theta_b)}{d\theta_b} = ((\alpha_a \frac{f_a^1(\eta(\theta_b))}{f_a^0(\eta(\theta_b))} - (1-\alpha_a))n_a(t) - (1-\alpha_b)n_b(t)) f_b^0(\theta_b) + \alpha_b f_b^1(\theta_b) n_b(t)$. Since $\frac{dL^t(\theta_b)}{d\theta_b}|_{\theta_b = \theta_b^{\max}} > 0$, based on $\frac{n_a(t)}{n_b(t)}$,

• If $\exists \theta_b'$ such that $\frac{dL^t(\theta_b)}{d\theta_b}|_{\theta_b = \theta_b'} \leq 0$, then one-shot decision $\theta_b(t)$ satisfies $\frac{n_a(t)}{n_b(t)} = \frac{1 - \frac{\alpha_b}{1-\alpha_b}\frac{f_b^1(\theta_b(t))}{f_b^0(\theta_b(t))}}{\frac{\alpha_a}{1-\alpha_a}\frac{f_a^1(\eta(\theta_b(t)))}{f_a^0(\eta(\theta_b(t)))} - 1}\frac{1-\alpha_b}{1-\alpha_a}$ and is unique.

• If $\frac{dL^t(\theta_b)}{d\theta_b} > 0, \forall \theta_b \in [\theta_b^{\min}, \theta_b^{\max}]$, then $\theta_b(t) < \theta_b^{\min}$ and $(\theta_a(t), \theta_b(t))$ does not satisfy Case 2.

2. $\mathcal{C} :=$ DP

$\frac{dL^t(\theta_b)}{d\theta_b} = n_b(t)(\alpha_b f_b^1(\theta_b) - (1-\alpha_b) f_b^0(\theta_b)) + n_a(t)((1-\alpha_b) f_b^0(\theta_b) + \alpha_b f_b^1(\theta_b))\frac{\alpha_a f_a^1(\eta(\theta_b)) - (1-\alpha_a) f_a^0(\eta(\theta_b))}{\alpha_a f_a^1(\eta(\theta_b)) + (1-\alpha_a) f_a^0(\eta(\theta_b))}$.

• If $\exists \theta_b(t)$ such that $\frac{dL^t(\theta_b)}{d\theta_b}|_{\theta_b = \theta_b(t)} = 0$, then it satisfies $\frac{n_a(t)}{n_b(t)} = (1 - \frac{2}{\frac{\alpha_b}{1-\alpha_b}\frac{f_b^1(\theta_b(t))}{f_b^0(\theta_b(t))} + 1})(\frac{2}{1 - \frac{\alpha_a f_a^1(\eta(\theta_b(t)))}{(1-\alpha_a) f_a^0(\eta(\theta_b(t)))}} - 1)$ and is unique.

• If $\frac{dL^t(\theta_b)}{d\theta_b} > 0, \forall \theta_b \in [\theta_b^{\min}, \theta_b^{\max}]$, then $\theta_b(t) < \theta_b^{\min}$ and $(\theta_a(t), \theta_b(t))$ does not satisfy Case 2.

• If $\frac{dL^t(\theta_b)}{d\theta_b} < 0, \forall \theta_b \in [\theta_b^{\min}, \theta_b^{\max}]$, then $\theta_b(t) > \theta_b^{\max}$ and $(\theta_a(t), \theta_b(t))$ does not satisfy Case 2.

Case 3: $\theta_a(t) \in [\underline{a}^1, \overline{a}^0]$, $\theta_b(t) \in [\overline{b}^0, \overline{b}^1]$

Express $L^t(\theta_a, \theta_b)$ as function of θ_a, the analysis will be similar to Case 1.

Let $\theta_a^{\min} = \max\{\underline{a}^1, \eta^{\mathcal{C}}(\overline{b}^0)\}$ be the minimum value θ_a can take.

$L^t(\theta_a) = n_a(t) \int_{\underline{a}^1}^{\theta_a} \alpha_a f_a^1(x) - (1-\alpha_a) f_a^0(x) dx + n_b(t) \int_{\underline{b}^1}^{(\eta^{\mathcal{C}})^{-1}(\theta_a)} \alpha_b f_b^1(x) dx + n_a(t) \int_{\underline{a}^1}^{\overline{a}^0} (1-\alpha_a) f_a^0(x) dx$

Taking derivative w.r.t. θ_a gives

$$\frac{dL^t(\theta_a)}{d\theta_a} = n_a(t)(\alpha_a f_a^1(\theta_a) - (1-\alpha_a) f_a^0(\theta_a)) + n_b(t) \alpha_b f_b^1((\eta^{\mathcal{C}})^{-1}(\theta_a)) \frac{d(\eta^{\mathcal{C}})^{-1}(\theta_a)}{d\theta_a},$$

where $\mathcal{C} :=$ `DP`.

$\frac{dL^t(\theta_a)}{d\theta_a} = n_a(t)(\alpha_a f_a^1(\theta_a) - (1-\alpha_a) f_a^0(\theta_a)) + n_b(t) \frac{\alpha_a f_a^1(\theta_a) + (1-\alpha_a) f_a^0(\theta_a)}{1 + \frac{(1-\alpha_b) f_b^0(\eta^{-1}(\theta_a))}{(1-\alpha_b) f_b^0(\eta^{-1}(\theta_a))}} = n_a(t)(\alpha_a f_a^1(\theta_a) - (1-\alpha_a) f_a^0(\theta_a)) + n_b(t)(\alpha_a f_a^1(\theta_a) + (1-\alpha_a) f_a^0(\theta_a))$, where the last equality holds since $f_b^0(\eta^{-1}(\theta_a)) = 0$ over $[\overline{b}^0, \overline{b}^1]$. Since $\frac{dL^t(\theta_b)}{d\theta_b}|_{\theta_b = \overline{a}^0} > 0$, based on the value of $\frac{n_a(t)}{n_b(t)}$,

• If $\exists \theta_a'$ such that $\frac{dL^t(\theta_a)}{d\theta_a}|_{\theta_a = \theta_a'} \le 0$, then one-shot decision $\theta_a(t)$ satisfies $\frac{n_a(t)}{n_b(t)} = \frac{2}{1 - \frac{\alpha_a f_a^1(\theta_a(t))}{(1-\alpha_a) f_a^0(\theta_a(t))}} - 1$ and is unique.

• If $\frac{dL^t(\theta_a)}{d\theta_a} > 0, \forall \theta_a \in [\underline{b}^1, \theta_a^{\min}]$, then $\theta_a(t) < \theta_a^{\min}$ and $(\theta_a(t), \theta_b(t))$ does not satisfy Case 3.

Now consider the case when $\mathcal{C} :=$ `Simple`, where $\theta_a(t) = \theta_b(t) = \theta(t)$. Since $\delta_a > \delta_b$, suppose that both $\delta_a, \delta_b \in \mathcal{T}_a \cap \mathcal{T}_b$ and according to Lemma 6, there could be only one case: $\theta(t) \in [\underline{a}^1, \overline{b}^0]$.

Taking derivative w.r.t. θ gives

$$\frac{dL^t(\theta)}{d\theta} = n_b(t)(\alpha_b f_b^1(\theta) - (1-\alpha_b) f_b^0(\theta)) + n_a(t)(\alpha_a f_a^1(\theta) - (1-\alpha_a) f_a^0(\theta)).$$

$\frac{dL^t(\theta)}{d\theta}$ is increasing from negative to positive over $[\delta_b, \delta_a]$, $\exists \theta(t)$ such that $\frac{dL^t(\theta)}{d\theta}|_{\theta=\theta(t)} = 0$, and it satisfies $\frac{n_a(t)}{n_b(t)} = \frac{\alpha_b f_b^1(\theta(t)) - (1-\alpha_b) f_b^0(\theta(t))}{(1-\alpha_a) f_a^0(\theta(t)) - \alpha_a f_a^1(\theta(t))}$.

□

By Lemma 6, $\theta_a(t) \in [\eta^{\mathcal{C}}(\delta_b), \delta_a], \theta_b(t) \in [\delta_b, (\eta^{\mathcal{C}})^{-1}(\delta_a)]$ hold. Under Assumption 7, $\alpha_b f_b^1(\theta_b) \ge (1-\alpha_b) f_b^0(\theta_b)$ for $\theta_b \in [\delta_b, \overline{b}^0]$, $\alpha_a f_a^1(\theta_a) \le (1-\alpha_a) f_a^0(\theta_a)$ for $\theta_a \in [\underline{a}^1, \delta_a]$. Moreover, $f_s^1(x)$ is

increasing and $f_s^0(x)$ is decreasing over $\mathcal{T}_s$. According to Lemma 16, for each case, function $\Psi_{\mathcal{C}}(\theta_a(t), \theta_b(t))$ is increasing in $\theta_a(t)$ and $\theta_b(t)$.

C.6 Proof of Theorem 15

Some notations are simplified by removing subscript t as mentioned in Appendix C.3.

Note that $f_{s,t}(x) = f_s(x)$ is fixed. Consider two one-shot problems under the same distributions at two consecutive time steps with group representation disparity $\frac{\widetilde{n}_a}{\widetilde{n}_b}$ and $\frac{\widehat{n}_a}{\widehat{n}_b}$ respectively. Let $(\widetilde{\theta}_a, \widetilde{\theta}_b)$ and $(\widehat{\theta}_a, \widehat{\theta}_b)$ be the corresponding solutions.

According to Lemma 6, $\widetilde{\theta}_a, \widehat{\theta}_a \in [\eta^{\mathcal{C}}(\delta_b), \delta_a]$, $\widetilde{\theta}_b, \widehat{\theta}_b \in [\delta_b, (\eta^{\mathcal{C}})^{-1}(\delta_a)]$ hold. Suppose $\frac{\widetilde{n}_a(t)}{\widetilde{n}_b(t)} > \frac{\widehat{n}_a}{\widehat{n}_b}$. By Theorem 14, it implies that $\widetilde{\theta}_a > \widehat{\theta}_a$, $\widetilde{\theta}_b > \widehat{\theta}_b$.

Consider the dynamics with $\lambda_s(\theta_s) = \nu(L_s(\theta_s))$, since $L_s(\theta_s)$ is decreasing over $[\underline{s}^0, \delta_s]$ and increasing over $[\delta_s, \overline{s}^1]$, the larger one-shot decisions θ_a, θ_b would result in the larger retention rate $\lambda_a(\theta_a)$ and the smaller $\lambda_b(\theta_b)$ as $\nu(\cdot)$ is strictly decreasing. Therefore, $\lambda_a(\widetilde{\theta}_a) > \lambda_a(\widehat{\theta}_a)$ and $\lambda_b(\widetilde{\theta}_b) < \lambda_b(\widehat{\theta}_b)$. Hence, Monotonicity condition is satisfied.

Consider the dynamics with $\lambda_s(\theta_s) = w(D_s(\theta_s))$ where $D_s(\theta_s) = \int_{\theta_s}^{\infty} \alpha_s f_s^1(x) - (1-\alpha_s) f_s^0(x) dx$. The following holds for $\mathcal{G}_a$ and $\mathcal{G}_b$:

$$D_a(\theta_a) = \int_{\delta_a}^{\infty} \alpha_a f_a^1(x) - (1-\alpha_a) f_a^0(x) dx + \int_{\theta_a}^{\delta_a} \alpha_a f_a^1(x) - (1-\alpha_a) f_a^0(x) dx$$
$$D_b(\theta_b) = \int_{\delta_b}^{\infty} g_b^1 f_b^1(x) - (1-\alpha_b) f_b^0(x) dx - \int_{\delta_b}^{\theta_b} g_b^1 f_b^1(x) - (1-\alpha_b) f_b^0(x) dx$$

Since $\alpha_s f_a^1(x) \le (1-\alpha_s) f_a^0(x)$ for $x \le \delta_a$ and $g_b^1 f_b^1(x) \ge (1-\alpha_b) f_b^0(x)$ for $x \ge \delta_b$, the larger θ_a, θ_b will thus result in the larger $\lambda_a(\theta_a)$ and smaller $\lambda_b(\theta_b)$ as $w(\cdot)$ is strictly increasing. Therefore, $\lambda_a(\widetilde{\theta}_a) > \lambda_a(\widehat{\theta}_a)$ and $\lambda_b(\widetilde{\theta}_b) < \lambda_b(\widehat{\theta}_b)$. Hence, Monotonicity condition is satisfied.

Combine with Theorem 12, $\frac{n_a(t)}{n_a(t)}$ changes monotonically. By Theorem 14, the corresponding one-shot fair decision $(\theta_a(t), \theta_b(t))$ also converges monotonically.

C.7 Proof of Theorem 16

C.7.1 Lemmas

To begin, we first introduce some lemmas for two cases. Lemma 17 and 19 show that under the same group representation n_a, n_b, the impact of reshaping distributions on the resulting one-shot decisions. Lemma 18 and 20 demonstrate a sufficient condition on feature distributions and one-shot decisions of two problems such that their expected losses satisfy certain conditions. The proof of these lemmas are presented in Appendix C.8.

Case (i): $f_{s,t}(x) = \alpha_{s,t} f_s^1(x) + (1-\alpha_{s,t}) f_s^0(x)$:

Fraction of subgroup $\mathcal{G}_s^y$ over $\mathcal{G}_s$ changes according to change of their own perceived loss L_s^y, i.e., for $i \in \{0,1\}$ such that $L_{s,t}^i(\theta_s(t)) < L_{s,t-1}^i(\theta_s(t-1))$, $\alpha_{s,t} > \alpha_{s,t-1}$ if $i = 1$ and $\alpha_{s,t} < \alpha_{s,t-1}$ if $i = 0$.

Lemma 17. *Let $(\widehat{\theta}_a, \widehat{\theta}_b)$, $(\widetilde{\theta}_a, \widetilde{\theta}_b)$ be two pairs of decisions under any of* `EqOpt`, `DP`, `Simple` *fairness criteria such that $\widehat{\Psi}_{\mathcal{C}}(\widehat{\theta}_a, \widehat{\theta}_b) = \widetilde{\Psi}_{\mathcal{C}}(\widetilde{\theta}_a, \widetilde{\theta}_b)$, where functions $\widehat{\Psi}_{\mathcal{C}}$, $\widetilde{\Psi}_{\mathcal{C}}$ have the form given in Table 4.1 and are defined under feature distributions $\widehat{f}_s(x) = \widehat{\alpha}_s f_s^1(x) + (1-\widehat{\alpha}_s) f_s^0(x)$, $\widetilde{f}_s(x) = \widetilde{\alpha}_s f_s^1(x) + (1-\widetilde{\alpha}_s) f_s^0(x)$ respectively $\forall s \in \{a,b\}$. If $\widehat{\alpha}_s < \widetilde{\alpha}_s$, then $\widehat{\theta}_s > \widetilde{\theta}_s$ will hold $\forall s \in \{a,b\}$.*

Lemma 18. *Consider two one-shot problems defined in* (4.1) *with objectives $\widetilde{\boldsymbol{O}}(\theta_a, \theta_b; \widetilde{n}_a, \widetilde{n}_b)$ and $\widehat{\boldsymbol{O}}(\theta_a, \theta_b; \widehat{n}_a, \widehat{n}_b)$, where $\widetilde{\boldsymbol{O}}$ is defined over distributions $\widetilde{f}_s(x) = (1-\widetilde{\alpha}_s) f_s^0(x) + \widetilde{\alpha}_s f_s^1(x)$ and $\widehat{\boldsymbol{O}}$ is defined over distributions $\widehat{f}_s(x) = (1-\widehat{\alpha}_s) f_s^0(x) + \widehat{\alpha}_s f_s^1(x)$, $s \in \{a,b\}$. Let $(\widetilde{\theta}_a, \widetilde{\theta}_b)$, $(\widehat{\theta}_a, \widehat{\theta}_b)$ be the corresponding one-shot decisions under any of* `Simple`, `EqOpt` *or* `DP` *fairness criteria. For any $\widehat{\alpha}, \widetilde{\alpha}_s)$ such that $\widehat{\alpha}_s < \widetilde{\alpha}_s$, $\forall s \in \{a,b\}$, if $\widehat{\theta}_a > \widetilde{\theta}_a$ and $\widehat{\theta}_b > \widetilde{\theta}_b$, then $\widehat{L}_a(\widehat{\theta}_a) < \widetilde{L}_a(\widetilde{\theta}_a)$ and $\widehat{L}_b(\widehat{\theta}_b) > \widetilde{L}_b(\widetilde{\theta}_b)$ can be satisfied under the following condition:*

$$\left|\Delta g_s (\widetilde{L}_s^0(\widetilde{\theta}_s) - \widetilde{L}_s^1(\widetilde{\theta}_s))\right| < \left| \int_{\widetilde{\theta}_s}^{\widehat{\theta}_s} (1-\widehat{\alpha}_s) f_s^0(x) - \widehat{\alpha}_s f_s^1(x) dx \right|, \ \forall s \in \{a,b\} \tag{C.3}$$

where $\Delta g_s = |\widehat{\alpha}_s - \widetilde{\alpha}_s|$.

Note that Condition (C.3) can be satisfied when: (1) Δg_s is sufficiently small; and (2) the difference in the decision $\widehat{\theta}_s - \widetilde{\theta}_s$ is sufficiently large, which can be achieved if $\widehat{n}_s$ and $\widetilde{n}_s$ are quite different.

Case (ii): $f_{s,t}(x) = \alpha_s f_{s,t}^1(x) + (1-\alpha_s) f_{s,t}^0(x)$

Suppose $L_{s,t}^1(\theta_s(t)) > L_{s,t-1}^1(\theta_s(t-1))$, i.e., $\mathcal{G}_s^1$ is less and less favored by the decision over time, then users from $\mathcal{G}_s^1$ will make additional effort to improve their features so that $f_{s,t}^1(x)$ will

skew toward the direction of higher feature value, i.e., $f^1_{s,t+1}(x) < f^1_{s,t}(x)$ for x with smaller value ($x \in \mathcal{T}_s$) while $\mathcal{G}^0_s$ is assumed to be unaffected, i.e., $f^0_{s,t+1}(x) = f^0_{s,t}(x)$. Similar statements hold when $\theta_s(t) < \theta_s(t-1)$ and $\mathcal{G}^0_s$ is less and less favored. Moreover, assume that Assumption 7 holds for any reshaped distributions and the support of $f^1_{s,t}(x)$ and $f^0_{s,t}(x)$ do not change over time.

$\forall t$, let $f^0_{s,t}(x)$ and $f^1_{s,t}(x)$ overlap over $\mathcal{T}_s := [\underline{s}^1, \overline{s}^0]$.

Lemma 19. *Let $(\widehat{\theta}_a, \widehat{\theta}_b)$, $(\widetilde{\theta}_a, \widetilde{\theta}_b)$ be two pairs of decisions under any of* `EqOpt`, `DP`, `Simple` *fairness criteria such that $\widehat{\Psi}_{\mathcal{C}}(\widehat{\theta}_a, \widehat{\theta}_b) = \widetilde{\Psi}_{\mathcal{C}}(\widetilde{\theta}_a, \widetilde{\theta}_b)$, where functions $\widehat{\Psi}_{\mathcal{C}}$, $\widetilde{\Psi}_{\mathcal{C}}$ have the form given in Table 4.1 and are defined under feature distributions $\widehat{f}_s(x) = \alpha_s \widehat{f}^1_s(x) + (1-\alpha_s)\widehat{f}^0_s(x)$, $\widetilde{f}_s(x) = \alpha_s \widetilde{f}^1_s(x) + (1-\alpha_s)\widetilde{f}^0_s(x)$ respectively $\forall s \in \{a,b\}$. If $\widehat{f}^0_s(x) = \widetilde{f}^0_s(x)$ and $\widehat{f}^1_s(x) < \widetilde{f}^1_s(x)$, $\forall x \in \mathcal{T}_s$, then $\widehat{\theta}_s > \widetilde{\theta}_s$ will hold $\forall s \in \{a,b\}$.*

Lemma 20. *Consider two one-shot problems defined in* (4.1) *with objectives $\widetilde{\boldsymbol{O}}(\theta_a, \theta_b; \widetilde{n}_a, \widetilde{n}_b)$ and $\widehat{\boldsymbol{O}}(\theta_a, \theta_b; \widehat{n}_a, \widehat{n}_b)$, where $\widetilde{\boldsymbol{O}}$ is defined over distributions $\widetilde{f}_s(x) = (1-\alpha_s)\widetilde{f}^0_s(x) + \alpha_s \widetilde{f}^1_s(x)$ and $\widehat{\boldsymbol{O}}$ is defined over distributions $\widehat{f}_s(x) = (1-\alpha_s)\widehat{f}^0_s(x) + \alpha_s \widehat{f}^1_s(x)$, $s \in \{a,b\}$. Let $(\widetilde{\theta}_a, \widetilde{\theta}_b)$, $(\widehat{\theta}_a, \widehat{\theta}_b)$ be the corresponding one-shot decisions under any of* `Simple`, `EqOpt` *or* `DP` *fairness criteria. For any distributions $\widetilde{f}^1_s$, $\widehat{f}^1_s$ increasing over $\mathcal{T}_s$ and $\widetilde{f}^0_s$, $\widehat{f}^0_s$ decreasing over $\mathcal{T}_s$ such that $\widehat{f}^1_s(x) < \widetilde{f}^1_s(x)$ over $\mathcal{T}_s$ and $\widehat{f}^0_s(x) = \widetilde{f}^0_s(x) = f^0_s(x), \forall x$, $\forall s \in \{a,b\}$. if $\widehat{\theta}_a > \widetilde{\theta}_a$ and $\widehat{\theta}_b > \widetilde{\theta}_b$, then $\widehat{L}_a(\widehat{\theta}_a) < \widetilde{L}_a(\widetilde{\theta}_a)$ holds. Moreover, $\widehat{L}_b(\widehat{\theta}_b) > \widetilde{L}_b(\widetilde{\theta}_b)$ can be satisfied under the following condition:*

$$\Delta f^1_b \alpha_b (\max\{\widetilde{\theta}_b, \widehat{\delta}_b\} - \underline{b}^1) < \int_{\max\{\widetilde{\theta}_b, \widehat{\delta}_b\}}^{\widehat{\theta}_b} \alpha_b \widehat{f}^1_b(x) - (1-\alpha_b)\widehat{f}^0_b(x) dx \tag{C.4}$$

where $\Delta f^1_b = \max_{x \in [\underline{b}^1, \max\{\widetilde{\theta}_b, \widehat{\delta}_b\}]} |\widehat{f}^1_b(x) - \widetilde{f}^1_b(x)|$ and $\widehat{\delta}_b$ is defined such that $g^0_b \widehat{f}^0_b(\widehat{\delta}_b) = g^1_b \widehat{f}^1_b(\widehat{\delta}_b)$.

Note that Condition (C.4) can be satisfied when: (1) Δf^1_b is sufficiently small, which makes $\widehat{\delta}_b$ close to $\widetilde{\delta}_b$ and $\widetilde{\theta}_b = \max\{\widetilde{\theta}_b, \widetilde{\delta}_b\}$ is more likely to hold; and (2) the difference in the decision $\widehat{\theta}_b - \widetilde{\theta}_b$ is sufficiently large, which can be achieved if $\widehat{n}_s$ and $\widetilde{n}_s$ are quite different.

C.7.2 Sufficient conditions

Below we formally state the sufficient condition under which Theorem 16 can hold.

Condition 3. *[Sufficient condition for exacerbation] Condition 3 is satisfied if the following holds:*

- *under* ***Case (i)****: Condition* (C.3) *is satisfied for objectives* $\boldsymbol{O}_t$ *and* $\boldsymbol{O}_{t+1}$*,* $\forall t \geq 2$*, i.e.,*

$$|\Delta g_{s,t+1}(L^0_{s,t}(\theta_s(t)) - L^1_{s,t}(\theta_s(t)))| < |\int_{\theta_s(t)}^{\theta_s(t+1)} (1-\alpha_{s,t+1}) f^0_s(x) - \alpha_{s,t+1} f^1_s(x) dx|, s \in \{a,b\}$$

with $\Delta g_{s,t+1} = |\alpha_{s,t+1} - \alpha_{s,t}|$.

- *under* ***Case (ii)****: Condition* (C.4) *is satisfied for objectives* $\boldsymbol{O}_t$ *and* $\boldsymbol{O}_{t+1}$*,* $\forall t \geq 2$*, i.e.,*

$$\Delta f^1_{b,t+1} \alpha_b (\max\{\theta_b(t), \delta_{b,t+1}\} - \underline{b}^1) < \int_{\max\{\theta_b(t), \delta_{b,t+1}\}}^{\theta_b(t+1)} \alpha_b f^1_{b,t+1}(x) - (1-\alpha_b) f^0_{b,t+1}(x) dx$$

with $\Delta f^1_{b,t+1} = \max_{x \in [\underline{b}^1, \max\{\theta_b(t), \delta_{b,t+1}\}]} |f^1_{b,t+1}(x) - f^1_{b,t}(x)|$.

Condition 4. *[Sufficient condition for acceleration of exacerbation]*

Let $\boldsymbol{O}^o_t := \boldsymbol{O}^o_t(\theta_a, \theta_b; n^o_a(t), n^o_b(t))$ *be the objective of the one-shot problem at time t for the case when distributions are fixed over time. Condition 4 is satisfied if the following holds:*

- *under* ***Case (i)****: Condition* (C.3) *is satisfied for objectives* $\boldsymbol{O}_t$ *and* $\boldsymbol{O}^o_t$*,* $\forall t \geq 2$*, i.e.,*

$$|\Delta g_{s,t}(L^0_{s,t}(\theta^o_s(t)) - L^1_{s,t}(\theta^o_s(t)))| < |\int_{\theta^o_s(t)}^{\theta_s(t)} (1-\alpha_{s,t}) f^0_s(x) - \alpha_{s,t} f^1_s(x) dx|, s \in \{a,b\}$$

with $\Delta g_{s,t} = \alpha_{s,t} - \alpha_{s,1}$.

- *under* ***Case (ii)****: Condition* (C.4) *is satisfied for objectives* $\boldsymbol{O}_t$ *and* $\boldsymbol{O}^o_t$*,* $\forall t \geq 2$*, i.e.,*

$$\Delta f^1_{b,t} \alpha_b (\max\{\theta^o_b(t), \delta_{b,t}\} - \underline{b}^1) < \int_{\max\{\theta^o_b(t), \delta_{b,t}\}}^{\theta_b(t)} \alpha_b f^1_{b,t}(x) - (1-\alpha_b) f^0_{b,t}(x) dx$$

with $\Delta f^1_{b,t} = \max_{x \in [\underline{b}^1, \max\{\theta^o_b(t), \delta_{b,t}\}]} |f^1_{b,t}(x) - f^1_{b,1}(x)|$.

Note that Condition 3 is likely to be satisfied when changing the decision from $\theta_s(t)$ to $\theta_s(t+1)$ results in: (i) a minor change of $f_{s,t+1}(x)$ from $f_{s,t}(x)$; or/and (ii) a significant change of representation disparity $\frac{n_a(t+1)}{n_b(t+1)}$ from $\frac{n_a(t)}{n_b(t)}$ so that $|\theta_s(t+1) - \theta_s(t)|$ is sufficiently large.

Condition 4 is likely to be satisfied if for any time step, (i) the change of $f_{s,t}(x)$ is minor as compared to the fixed distribution, i.e., $f_{s,1}(x)$ at time $t = 1$; or/and (ii) the resulting decisions at same time under two schemes are quite different, i.e., $|\theta^o_s(t) - \theta_s(t)|$ is sufficiently large.

In other words, both requires that $f_{s,t}(x)$ is relatively insensitive to the change of one-shot decisions, and this applies to scenarios where the impact of reshaping distributions is considered as a slow process, e.g., change of credit score takes time and is a slow process.

C.7.3 Proof of main theorem

If $f_{s,t}(x) = f_s(x)$ is fixed $\forall t$, then the relationship between $\frac{n_a^o(t)}{n_b^o(t)}$ and one-shot solutions $(\theta_a^o(t), \theta_b^o(t))$ follows $\frac{n_a^o(t)}{n_b^o(t)} = \Psi_{\mathcal{C},1}(\theta_a^o(t), \theta_b^o(t)), \forall t$. If $f_{s,t}(x)$ varies over time, then $\frac{n_a(t)}{n_b(t)} = \Psi_{\mathcal{C},t}(\theta_a(t), \theta_b(t)), \forall t$. We consider that distributions start to change after individuals feel the change of perceived decisions, i.e., $f_{s,t}(x)$ begins to change at time $t = 3$. In the following, $\forall s \in \{a, b\}$, $\theta_s^o(t) = \theta_s(t)$, $\lambda_{s,t}^o(\theta_s^o(t)) = \lambda_{s,t}(\theta_s(t))$ for $t = 1, 2$ and $\frac{n_a^o(t)}{n_b^o(t)} = \frac{n_a(t)}{n_b(t)}$ for $t = 1, 2, 3$.

Start from $t = 1$, if $(\theta_a(1), \theta_b(1))$ satisfies $\lambda_{a,1}(\theta_a(1)) > \lambda_{b,1}(\theta_b(1))$, then $\frac{n_a(2)}{n_b(2)} > \frac{n_a(1)}{n_b(1)}$ and $\theta_s(2) > \theta_s(1)$ holds $\forall s \in \{a, b\}$, implying $\lambda_{a,2}(\theta_a(2)) > \lambda_{a,1}(\theta_a(1)) > \lambda_{b,1}(\theta_b(1)) > \lambda_{b,2}(\theta_b(2))$ ($\boldsymbol{O}_1$ *and* $\boldsymbol{O}_2$ *satisfy monotonicity condition*) and $\frac{n_a(3)}{n_b(3)} > \frac{n_a(2)}{n_b(2)}$. Moreover, the change of decisions begins to reshape the feature distributions in the next time step.

Consider two ways of reshaping distributions: **Case (i)** and **Case (ii)**. For both cases, show that as long as the change of distribution from $f_{s,t-1}(x)$ to $f_{s,t}(x)$ is relatively small w.r.t. the change of decision from $\theta_s(t-2)$ to $\theta_s(t-1)$ (formally stated in Condition 3 and Condition 4), the following can hold for any time step $t \geq 3$: *(i)* $\boldsymbol{O}_t$ and $\boldsymbol{O}_{t+1}$ satisfy monotonicity condition: $\lambda_{a,t+1}(\theta_a(t+1)) > \lambda_{a,t}(\theta_a(t))$, $\lambda_{b,t}(\theta_b(t)) > \lambda_{b,t+1}(\theta_b(t+1))$ hold when $\frac{n_a(t+1)}{n_b(t+1)} > \frac{n_a(t)}{n_b(t)}$; *(ii)* group representation disparity changes faster than case when distributions are fixed, i.e., $\frac{n_a(t)}{n_b(t)} \geq \frac{n_a^o(t)}{n_b^o(t)}, \forall t$.

Since $\theta_s(2) > \theta_s(1)$, within the same group $\mathcal{G}_s$, subgroup $\mathcal{G}_s^1$ (resp. $\mathcal{G}_s^0$) experiences the higher (resp. lower) loss at time $t = 2$ than $t = 1$. Consider two types of change $\forall s \in \{a, b\}$:

- **Case (i)**: $\alpha_{s,3} < \alpha_{s,2} = \alpha_{s,1}$.
- **Case (ii)**: $f_{s,3}^0(x) = f_{s,2}^0(x) = f_{s,1}^0(x), \forall x$ and $f_{s,3}^1(x) < f_{s,2}^1(x) = f_{s,1}^1(x), \forall x \in \mathcal{T}_s$.

Prove the following by induction under Condition 3 and 4 (on the sensitivity of $f_{s,t}(x)$ w.r.t. the change of decisions): For $t > 3$, $\frac{n_a(t+1)}{n_b(t+1)} > \frac{n_a^o(t+1)}{n_b^o(t+1)}$ and $\frac{n_a(t+1)}{n_b(t+1)} > \frac{n_a(t)}{n_b(t)}$ hold, and $\forall s \in \{a, b\}$:

- **Case (i)**: $\alpha_{s,t+1} < \alpha_{s,t} < \alpha_{s,1}$ is satisfied.
- **Case (ii)**: $f_{s,t+1}^0(x) = f_{s,t}^0(x) = f_{s,1}^0(x), \forall x$ and $f_{s,t+1}^1(x) < f_{s,t}^1(x) < f_{s,1}^1(x), \forall x \in \mathcal{T}_s$.

Base case:

$\Psi_{\mathcal{C},t}$ is defined under feature distributions $f_{s,t}(x) = \alpha_{s,t} f_{s,t}^1(x) + (1 - \alpha_{s,t}) f_{s,t}^0(x)$, $\forall s \in \{a, b\}$. Define a pair $(\tilde{\theta}_a, \tilde{\theta}_b)$ such that the following holds:

$\frac{n_a(3)}{n_b(3)} = \Psi_{\mathcal{C},1}(\theta_a^o(3), \theta_b^o(3)) = \Psi_{\mathcal{C},3}(\theta_a(3), \theta_b(3)) = \Psi_{\mathcal{C},2}(\tilde{\theta}_a, \tilde{\theta}_b) > \Psi_{\mathcal{C},2}(\theta_a(2), \theta_b(2)) = \frac{n_a(2)}{n_b(2)}$.

Then, we have $\forall s \in \{a, b\}$:

• **Case (i)**: As $\alpha_{s,3} < \alpha_{s,2}^1 = \alpha_{s,1}$, by Lemma 17, $\theta_s(3) > \theta_s^o(3) = \tilde{\theta}_s$ holds.

• **Case (ii)**: As $f_{s,3}^0(x) = f_{s,2}^0(x) = f_{s,1}^0(x), \forall x$ and $f_{s,3}^1(x) < f_{s,2}^1(x) < f_{s,1}^1(x), \forall x \in \mathcal{T}_s$, by Lemma 19, $\theta_s(3) > \theta_s^o(3) = \tilde{\theta}_s$ holds.

By Theorem 14, $\tilde{\theta}_s > \theta_s(2)$ holds. It implies that $\theta_s(3) > \theta_s^o(3)$ and $\theta_s(3) > \theta_s(2)$.

Consider dynamics with $\lambda_{s,t}(\theta_s(t)) = \nu(L_{s,t}(\theta_s(t)))$. The following statements hold:

(1) Under Condition 3, $L_{a,3}(\theta_a(3)) < L_{a,2}(\theta_a(2))$ and $L_{b,3}(\theta_b(3)) > L_{b,2}(\theta_b(2))$ hold, implying $\lambda_{a,3}(\theta_a(3)) > \lambda_{a,2}(\theta_a(2)) > \lambda_{b,2}(\theta_b(2)) > \lambda_{b,3}(\theta_b(3))$ and $\frac{n_a(4)}{n_b(4)} > \frac{n_a(3)}{n_b(3)}$.

(2) Under Condition 4, $L_{a,3}(\theta_a(3)) < L_{a,3}(\theta_a^o(3))$ and $L_{b,3}(\theta_b(3)) > L_{b,3}(\theta_b^o(3))$ hold, implying $\lambda_{a,3}(\theta_a(3)) > \lambda_{a,3}^o(\theta_a^o(3)) > \lambda_{b,3}^o(\theta_b^o(3)) > \lambda_{b,3}(\theta_b(3))$ and $\frac{n_a(4)}{n_b(4)} > \frac{n_a^o(4)}{n_b^o(4)}$.

(3) $\mathcal{G}_s^1$ (resp. $\mathcal{G}_s^0$) experiences the higher (resp. lower) loss at $t = 3$ than $t = 2$, i.e., $L_{s,3}^1(\theta_s(3)) > L_{s,2}^1(\theta_s(2))$ and $L_{s,3}^0(\theta_s(3)) < L_{s,2}^0(\theta_s(2))$,

• **Case (i)**: $\alpha_{s,4} < \alpha_{s,3} < \alpha_{s,1}$ holds.

• **Case (ii)**: $f_{s,4}^0(x) = f_{s,3}^0(x) = f_{s,1}^0(x), \forall x$ and $f_{s,4}^1(x) < f_{s,3}^1(x) < f_{s,1}^1(x), \forall x \in \mathcal{T}_s$ hold.

Induction step:

Suppose at time $t > 3$, $\frac{n_a(t+1)}{n_b(t+1)} > \frac{n_a^o(t+1)}{n_b^o(t+1)}$ and $\frac{n_a(t+1)}{n_b(t+1)} > \frac{n_a(t)}{n_b(t)}$ hold, and $\forall s \in \{a, b\}$:

• **Case (i)**: $\alpha_{s,t+1} < \alpha_{s,t} < \alpha_{s,1}$ is satisfied.

• **Case (ii)**: $f_{s,t+1}^0(x) = f_{s,t}^0(x) = f_{s,1}^0(x), \forall x$ and $f_{s,t+1}^1(x) < f_{s,t}^1(x) < f_{s,1}^1(x), \forall x \in \mathcal{T}_s$.

Then consider time step $t + 1$.

Define pairs $(\tilde{\theta}_a, \tilde{\theta}_b)$ and $(\hat{\theta}_a, \hat{\theta}_b)$ such that the following holds:

$$\frac{n_a(t+1)}{n_b(t+1)} = \Psi_{\mathcal{C},t+1}(\theta_a(t+1), \theta_b(t+1)) > \begin{cases} \frac{n_a^o(t+1)}{n_b^o(t+1)} = \Psi_{\mathcal{C},1}(\theta_a^o(t+1), \theta_b^o(t+1)) = \Psi_{\mathcal{C},t+1}(\tilde{\theta}_a, \tilde{\theta}_b) \\ \frac{n_a(t)}{n_b(t)} = \Psi_{\mathcal{C},t}(\theta_a(t), \theta_b(t)) = \Psi_{\mathcal{C},t+1}(\hat{\theta}_a, \hat{\theta}_b) \end{cases}$$

According to the hypothesis, Under **Case (i)**, $\tilde{\theta}_s > \theta_s^o(t+1)$ and $\hat{\theta}_s > \theta_s(t)$ hold by Lemma 17. Under **Case (ii)**, $\tilde{\theta}_s > \theta_s^o(t+1)$ and $\hat{\theta}_s > \theta_s(t)$ hold by Lemma 19. By Theorem 14, $\theta_s(t+1) > \tilde{\theta}_s$ and $\theta_s(t+1) > \hat{\theta}_s$ hold. It implies that $\theta_s(t+1) > \theta_s^o(t+1)$ and $\theta_s(t+1) > \theta_s(t)$.

(1) Under Condition 3, $L_{a,t+1}(\theta_a(t+1)) < L_{a,t}(\theta_a(t))$ and $L_{b,t+1}(\theta_b(t+1)) > L_{b,t}(\theta_b(t))$hold, implying $\lambda_{a,t+1}(\theta_a(t+1)) > \lambda_{a,t}(\theta_a(t)) > \lambda_{b,t}(\theta_b(t)) > \lambda_{b,t+1}(\theta_b(t+1))$ and $\frac{n_a(t+1)}{n_b(t+1)} > \frac{n_a(t)}{n_b(t)}$: $\boldsymbol{O}_t$ and $\boldsymbol{O}_{t+1}$ satisfy monotonicity condition and representation disparity get exacerbated.

(2) Under Condition 4, $L_{a,t+1}(\theta_a(t+1)) < L_{a,t+1}(\theta_a^o(t+1))$ and $L_{b,t+1}(\theta_b(t+1)) > L_{b,t+1}(\theta_b^o(t+1))$

hold, implying $\lambda_{a,t+1}(\theta_a(t+1)) > \lambda^o_{a,t+1}(\theta^o_a(t+1)) > \lambda^o_{b,t+1}(\theta^o_b(t+1)) > \lambda_{b,t+1}(\theta_b(t+1))$ and thus $\frac{n_a(t+1)}{n_b(t+1)} > \frac{n^o_a(t+1)}{n^o_b(t+1)}$: the discrepancy between retention rates of two demographic groups is larger at each time compared to the case when distributions are fixed, and if the disparity get exacerbated, this exacerbation is accelerated under the reshaping.

(3) $\mathcal{G}^1_s$ (resp. $\mathcal{G}^0_s$) experiences the higher (resp. lower) loss at $t+1$ than t, i.e., $L^1_{s,t+1}(\theta_s(t+1)) > L^1_{s,t}(\theta_s(t))$ and $L^0_{s,t+1}(\theta_s(t+1)) < L^0_{s,t}(\theta_s(t))$. Therefore,

• **Case (i)**: $\alpha_{s,t+2} < \alpha_{s,t+1} < \alpha_{s,1}$ holds.

• **Case (ii)**: $f^0_{s,t+2}(x) = f^0_{s,t+1}(x) = f^0_{s,1}(x), \forall x$ and $f^1_{s,t+2}(x) < f^1_{s,t+1}(x) < f^1_{s,1}(x), \forall x \in \mathcal{T}_s$ hold.

Proof is completed.

The case if $\lambda_{a,1}(\theta_a(1)) < \lambda_{b,1}(\theta_b(1))$ can be proved similarly and is omitted.

C.8 Proof of Lemmas for Theorem 16

C.8.1 Proof of Lemma 17

$f^0_s(x)$ and $f^1_s(x)$ overlap over $\mathcal{T}_s := [\underline{s}^1, \overline{s}^0]$.

1. $\mathcal{C} :=$ `DP`

To satisfy $\widehat{\Psi}_{\texttt{DP}}(\widehat{\theta}_a, \widehat{\theta}_b) = \widetilde{\Psi}_{\texttt{DP}}(\widetilde{\theta}_a, \widetilde{\theta}_b)$, $\frac{\widehat{\alpha}_s f^1_s(\widehat{\theta}_s)}{(1-\widehat{\alpha}_s) f^0_s(\widehat{\theta}_s)} = \frac{\widetilde{\alpha}_s f^1_s(\widetilde{\theta}_s)}{(1-\widetilde{\alpha}_s) f^0_s(\widetilde{\theta}_s)}$ should hold. Under Assumption 7, both $\frac{\widehat{\alpha}_s f^1_s(\cdot)}{(1-\widehat{\alpha}_s) f^0_s(\cdot)}$ and $\frac{\widetilde{\alpha}_s f^1_s(\cdot)}{(1-\widetilde{\alpha}_s) f^0_s(\cdot)}$ are strictly increasing over $\mathcal{T}_s$. Since $\forall s \in \{a,b\}$, there is $\frac{\widehat{\alpha}_s f^1_s(\theta_s)}{(1-\widehat{\alpha}_s) f^0_s(\theta_s)} < \frac{\widetilde{\alpha}_s f^1_s(\theta_s)}{(1-\widetilde{\alpha}_s) f^0_s(\theta_s)}$, $\forall \theta_s \in \mathcal{T}_s$. For all three possibilities in Table 4.1, $\widehat{\theta}_s > \widetilde{\theta}_s$ holds $\forall s \in \{a,b\}$.

2. $\mathcal{C} :=$ `EqOpt`

Since $\widetilde{L}^0_a(\theta_a) = \widetilde{L}^0_b(\theta_b)$ and $\widehat{L}^0_a(\theta_a) = \widehat{L}^0_b(\theta_b)$ always hold for any (θ_a, θ_b) satisfying `EqOpt` criterion, when change of $\widehat{\alpha}_s$ (or $\widetilde{\alpha}_s$) is determined by θ_s only via $\widehat{L}^0_s(\theta_s)$ (or $\widetilde{L}^0_s(\theta_s)$), both $\frac{\widehat{\alpha}_b}{\widehat{\alpha}_a} = 1$ and $\frac{\widetilde{\alpha}_b}{\widetilde{\alpha}_a} = 1$ are satisfied. To satisfy $\widehat{\Psi}_{\texttt{EqOpt}}(\widehat{\theta}_a, \widehat{\theta}_b) = \widetilde{\Psi}_{\texttt{EqOpt}}(\widetilde{\theta}_a, \widetilde{\theta}_b)$, $\frac{\widehat{\alpha}_s f^1_s(\widehat{\theta}_s)}{(1-\widehat{\alpha}_s) f^0_s(\widehat{\theta}_s)} = \frac{\widetilde{\alpha}_s f^1_s(\widetilde{\theta}_s)}{(1-\widetilde{\alpha}_s) f^0_s(\widetilde{\theta}_s)}$ should hold, which is same as the condition that should be satisfied in case when $\mathcal{C} :=$ `DP`. Rest of the proof is thus same as `DP` case and is omitted.

3. $\mathcal{C} :=$ `Simple`

`Simple` fairness criterion requires that $\widehat{\theta}_a = \widehat{\theta}_b = \widehat{\theta}$ and $\widetilde{\theta}_a = \widetilde{\theta}_b = \widetilde{\theta}$. In order to satisfy $\widehat{\Psi}_{\texttt{Simple}}(\widehat{\theta}_a, \widehat{\theta}_b) = \widetilde{\Psi}_{\texttt{Simple}}(\widetilde{\theta}_a, \widetilde{\theta}_b)$, $\frac{\widehat{\alpha}_b f^1_b(\widehat{\theta}) - (1-\widehat{\alpha}_b) f^0_b(\widehat{\theta})}{(1-\widehat{\alpha}_a) f^0_a(\widehat{\theta}) - \widehat{\alpha}_a f^1_a(\widehat{\theta})} = \frac{\widetilde{\alpha}_b f^1_b(\widetilde{\theta}) - (1-\widetilde{\alpha}_b) f^0_b(\widetilde{\theta})}{(1-\widetilde{\alpha}_a) f^0_a(\widetilde{\theta}) - \widetilde{\alpha}_a f^1_a(\widetilde{\theta})}$ should hold. Under Assumption 7, both $\frac{\widehat{\alpha}_b f^1_b(\cdot) - (1-\widehat{\alpha}_b) f^0_b(\cdot)}{(1-\widehat{\alpha}_a) f^0_a(\cdot) - \widehat{\alpha}_a f^1_a(\cdot)}$ and $\frac{\widetilde{\alpha}_b f^1_b(\cdot) - (1-\widetilde{\alpha}_b) f^0_b(\cdot)}{(1-\widetilde{\alpha}_a) f^0_a(\cdot) - \widetilde{\alpha}_a f^1_a(\cdot)}$ are strictly increasing over $\mathcal{T}_s$. Since

$\forall s \in \{a,b\}$, there is $\frac{\widehat{\alpha}_b f_b^1(\theta)-(1-\widehat{\alpha}_b)f_b^0(\theta)}{(1-\widehat{\alpha}_a)f_a^0(\theta)-\widehat{\alpha}_a f_a^1(\theta)} < \frac{\widetilde{\alpha}_b f_b^1(\theta)-(1-\widetilde{\alpha}_b)f_b^0(\theta)}{(1-\widetilde{\alpha}_a)f_a^0(\theta)-\widetilde{\alpha}_a f_a^1(\theta)}$, $\forall \theta \in \mathcal{T}_a \cap \mathcal{T}_b$, implying that $\widehat{\theta} > \widetilde{\theta}$.

C.8.2 Proof of Lemma 18

Define $\Delta L_s^j = |\widehat{L}_s^j(\widehat{\theta}_s) - \widetilde{L}_s^j(\widetilde{\theta}_s)|$, $j \in \{0,1\}$. Rewrite $\widehat{\alpha}_s = \widetilde{\alpha}_s - \Delta g_s$. For $s \in \{a,b\}$, $\widehat{\theta}_s > \widetilde{\theta}_s$ holds, which implies that $\widehat{L}_s^1(\widehat{\theta}_s) = \widetilde{L}_s^1(\widetilde{\theta}_s) + \Delta L_s^1$ and $\widehat{L}_s^0(\widehat{\theta}_s) = \widetilde{L}_s^0(\widetilde{\theta}_s) - \Delta L_s^0$. Therefore,

$$\widehat{L}_s(\widehat{\theta}_s) - \widetilde{L}_s(\widetilde{\theta}_s) = \Delta g_s(\widetilde{L}_s^0(\widetilde{\theta}_s) - \widetilde{L}_s^1(\widetilde{\theta}_s)) - ((1-\widehat{\alpha}_s)\Delta L_s^0 - \widehat{\alpha}_s \Delta L_s^1), s \in \{a,b\}$$

since

$$\Delta L_s^1 = \int_{\widetilde{\theta}_s}^{\widehat{\theta}_s} f_s^1(x)dx;\ \Delta L_s^0 = \int_{\widetilde{\theta}_s}^{\widehat{\theta}_s} f_s^0(x)dx$$

Define $\widehat{\delta}_s$ such that $(1-\widehat{\alpha}_s)f_s^0(\widehat{\delta}_s) = \widehat{\alpha}_s f_s^1(\widehat{\delta}_s)$, then $(1-\widehat{\alpha}_a)f_a^0(x) > \widehat{\alpha}_a f_a^1(x)$ when $x < \widehat{\delta}_a$ and $(1-\widehat{\alpha}_b)f_b^0(x) < \widehat{\alpha}_b f_b^1(x)$ when $x > \widehat{\delta}_b$. By Lemma 6, $\widehat{\theta}_a < \widehat{\delta}_a$ and $\widehat{\theta}_b > \widehat{\delta}_b$ hold, implying

$$(1-\widehat{\alpha}_s)\Delta L_s^0 - \widehat{\alpha}_s \Delta L_s^1 = \int_{\widetilde{\theta}_s}^{\widehat{\theta}_s} (1-\widehat{\alpha}_s)f_s^0(x) - \widehat{\alpha}_s f_s^1(x)dx \begin{cases} > 0, & s = a \\ < 0, & s = b \end{cases}$$

If $|\Delta g_s(\widetilde{L}_s^0(\widetilde{\theta}_s) - \widetilde{L}_s^1(\widetilde{\theta}_s))| < |\int_{\widetilde{\theta}_s}^{\widehat{\theta}_s}(1-\widehat{\alpha}_s)f_s^0(x) - \widehat{\alpha}_s f_s^1(x)dx|$ holds, then the sign of $\widehat{L}_s(\widehat{\theta}_s) - \widetilde{L}_s(\widetilde{\theta}_s)$ is determined by the sign of $\widehat{\alpha}_s \Delta L_s^1 - (1-\widehat{\alpha}_s)\Delta L_s^0$. We have $\widehat{L}_a(\widehat{\theta}_a) < \widetilde{L}_a(\widetilde{\theta}_a)$ and $\widehat{L}_b(\widehat{\theta}_b) > \widetilde{L}_b(\widetilde{\theta}_b)$.

C.8.3 Proof of Lemma 19

1. $\mathcal{C} :=$ `DP` or $\mathcal{C} :=$ `EqOpt`

To satisfy $\widehat{\Psi}_{\texttt{DP}}(\widehat{\theta}_a,\widehat{\theta}_b) = \widetilde{\Psi}_{\texttt{DP}}(\widetilde{\theta}_a,\widetilde{\theta}_b)$ or $\widehat{\Psi}_{\texttt{EqOpt}}(\widehat{\theta}_a,\widehat{\theta}_b) = \widetilde{\Psi}_{\texttt{EqOpt}}(\widetilde{\theta}_a,\widetilde{\theta}_b)$, $\frac{\alpha_s \widehat{f}_s^1(\widetilde{\theta}_s)}{(1-\alpha_s)\widehat{f}_s^0(\widetilde{\theta}_s)} = \frac{\alpha_s \widehat{f}_s^1(\widehat{\theta}_s)}{(1-\alpha_s)\widehat{f}_s^0(\widehat{\theta}_s)} < \frac{\alpha_s \widetilde{f}_s^1(\widehat{\theta}_s)}{(1-\alpha_s)\widetilde{f}_s^0(\widehat{\theta}_s)}$ should hold. Under Assumption 7, $\frac{\alpha_s \widetilde{f}_s^1(\cdot)}{(1-\alpha_s)\widetilde{f}_s^0(\cdot)}$ is strictly increasing over $\mathcal{T}_s$. $\widehat{\theta}_s > \widetilde{\theta}_s$ has to be satisfied.

2. $\mathcal{C} :=$ `Simple`

`Simple` fairness criterion requires that $\widehat{\theta}_a = \widehat{\theta}_b = \widehat{\theta}$ and $\widetilde{\theta}_a = \widetilde{\theta}_b = \widetilde{\theta}$. In order to satisfy $\widehat{\Psi}_{\texttt{Simple}}(\widehat{\theta}_a,\widehat{\theta}_b) = \widetilde{\Psi}_{\texttt{Simple}}(\widetilde{\theta}_a,\widetilde{\theta}_b)$, $\frac{g_b^1 \widehat{f}_b^1(\widetilde{\theta}) - g_b^0 \widehat{f}_b^0(\widetilde{\theta})}{g_a^0 \widehat{f}_a^0(\widetilde{\theta}) - g_a^1 \widehat{f}_a^1(\widetilde{\theta})} = \frac{g_b^1 \widehat{f}_b^1(\widehat{\theta}) - g_b^0 \widehat{f}_b^0(\widehat{\theta})}{g_a^0 \widehat{f}_a^0(\widehat{\theta}) - g_a^1 \widehat{f}_a^1(\widehat{\theta})} < \frac{g_b^1 \widetilde{f}_b^1(\widehat{\theta}) - g_b^0 \widetilde{f}_b^0(\widehat{\theta})}{g_a^0 \widetilde{f}_a^0(\widehat{\theta}) - g_a^1 \widetilde{f}_a^1(\widehat{\theta})}$ should hold. Un-

der Assumption 7, $\frac{g_b^1 \widehat{f}_b^1(\cdot) - g_b^0 \widehat{f}_b^0(\cdot)}{g_a^0 \widehat{f}_a^0(\cdot) - g_a^1 \widehat{f}_a^1(\cdot)}$ is strictly increasing over $\mathcal{T}_s$. For $\widehat{\theta}, \widetilde{\theta} \in \mathcal{T}_a \cap \mathcal{T}_b$, $\widehat{\theta} > \widetilde{\theta}$ has to be satisfied.

C.8.4 Proof of Lemma 20

Define $\widehat{\delta}_s$ such that $(1-\alpha_s)\widehat{f}_s^0(\widehat{\delta}_s) = \alpha_s \widehat{f}_s^1(\widehat{\delta}_s)$. Then, $(1-\alpha_a)\widehat{f}_a^0(x) > \alpha_a^1 \widehat{f}_a^1(x)$ when $x < \widehat{\delta}_a$ and $(1-\alpha_b)\widehat{f}_b^0(x) < \alpha_b \widehat{f}_b^1(x)$ when $x > \widehat{\delta}_b$.

Since $\widehat{\theta}_s > \widetilde{\theta}_s$, we have

$$\begin{aligned}
\widehat{L}_s^0(\widehat{\theta}_s) - \widetilde{L}_s^0(\widetilde{\theta}_s) &= -\int_{\widetilde{\theta}_s}^{\widehat{\theta}_s} \widehat{f}_s^0(x)dx = -\int_{\widetilde{\theta}_s}^{\widehat{\theta}_s} \widehat{f}_s^0(x)dx \\
\widehat{L}_s^1(\widehat{\theta}_s) - \widetilde{L}_s^1(\widetilde{\theta}_s) &= \int_{\widetilde{\theta}_s}^{\widehat{\theta}_s} \widehat{f}_s^1(x)dx - \int_{\underline{s}^1}^{\widetilde{\theta}_s} (\widehat{f}_s^1(x) - \widehat{f}_s^1(x))dx
\end{aligned}$$

Therefore,

$$\widehat{L}_s(\widehat{\theta}_s) - \widetilde{L}_s(\widetilde{\theta}_s) = \int_{\widetilde{\theta}_s}^{\widehat{\theta}_s} \alpha_s \widehat{f}_s^1(x) - (1-\alpha_s)\widehat{f}_s^0(x)dx - \alpha_s \int_{\underline{s}^1}^{\widetilde{\theta}_s} (\widehat{f}_s^1(x) - \widehat{f}_s^1(x))dx$$

since $\widetilde{\theta}_a < \widehat{\theta}_a < \widehat{\delta}_a$, $\int_{\widetilde{\theta}_a}^{\widehat{\theta}_a} \alpha_a \widehat{f}_a^1(x) - g_a^0 \widehat{f}_a^0(x)dx < 0$ holds. Since $\widetilde{f}_a^1(x) > \widehat{f}_a^1(x)$ for $x \in \mathcal{T}_a$, we have $\alpha_a \int_{\underline{a}^1}^{\widetilde{\theta}_a} (\widetilde{f}_a^1(x) - \widehat{f}_a^1(x))dx > 0$. Therefore, $\widehat{L}_a(\widehat{\theta}_a) < \widetilde{L}_a(\widetilde{\theta}_a)$.

When $s = b$, there are two possibilities: (i) $\widetilde{\theta}_b < \widehat{\delta}_b < \widehat{\theta}_b$; (ii) $\widehat{\delta}_b < \widetilde{\theta}_b < \widehat{\theta}_b$.

For case (i),

$$\begin{aligned}
\widehat{L}_b(\widehat{\theta}_b) - \widetilde{L}_b(\widetilde{\theta}_b) &= \underbrace{\int_{\widehat{\delta}_b}^{\widehat{\theta}_b} \alpha_b \widehat{f}_b^1(x) - (1-\alpha_b)\widehat{f}_b^0(x)dx}_{\textbf{term 1}} + \underbrace{\int_{\widetilde{\theta}_b}^{\widehat{\delta}_b} \alpha_b \widehat{f}_b^1(x) - (1-\alpha_b)\widehat{f}_b^0(x)dx}_{\textbf{term 2}} \\
&+ \underbrace{\alpha_b \int_{\underline{b}^1}^{\widetilde{\theta}_b} (\widehat{f}_b^1(x) - \widetilde{f}_b^1(x))dx}_{\textbf{term 3}}
\end{aligned}$$

Since $\widetilde{\delta}_b < \widetilde{\theta}_b < \widehat{\delta}_b$ and $\widehat{f}_b^0(x) = \widetilde{f}_b^0(x)$, for $x \in [\widetilde{\theta}_b, \widehat{\delta}_b]$, $\alpha_b \widehat{f}_b^1(x) - \alpha_b \widetilde{f}_b^1(x) < \alpha_b \widehat{f}_b^1(x) - (1-\alpha_b)\widehat{f}_b^0(x) < 0$, we have $0 >$ **term 2** + **term 3** $> \alpha_b \int_{\underline{b}^1}^{\widehat{\delta}_b} (\widehat{f}_b^1(x) - \widetilde{f}_b^1(x))dx$.

Define $\Delta_1 = \max_{x\in[\underline{b}^1,\widehat{\delta}_b]} |\widehat{f}_b^1(x) - \widetilde{f}_b^1(x)|$. Since **term 1** > 0, $\widehat{L}_b(\widehat{\theta}_b) > \widetilde{L}_b(\widetilde{\theta}_b)$ holds only if the following condition is satisfied:

$$\Delta_1 \alpha_b (\widehat{\delta}_b - \underline{b}^1) < \int_{\widehat{\delta}_b}^{\widehat{\theta}_b} \alpha_b \widehat{f}_b^1(x) - (1-\alpha_b)\widehat{f}_b^0(x)dx$$

For case (ii),

$$\widehat{L}_b(\widehat{\theta}_b) - \widetilde{L}_b(\widetilde{\theta}_b) = \underbrace{\int_{\widetilde{\theta}_b}^{\widehat{\theta}_b} \alpha_b \widehat{f}_b^1(x) - (1-\alpha_b)\widehat{f}_b^0(x)dx}_{\textbf{term 1}} + \underbrace{\alpha_b \int_{\underline{b}^1}^{\widetilde{\theta}_b} (\widehat{f}_b^1(x) - \widetilde{f}_b^1(x))dx}_{\textbf{term 2}}$$

Define $\Delta_2 = \max_{x\in[\underline{b}^1,\widetilde{\theta}_b]} |\widehat{f}_b^1(x) - \widetilde{f}_b^1(x)|$. Similar to case (i), $\widehat{L}_b(\widehat{\theta}_b) > \widetilde{L}_b(\widetilde{\theta}_b)$ holds only if the following condition is satisfied:

$$\Delta_2 \alpha_b (\widetilde{\theta}_b - \underline{b}^1) < \int_{\widetilde{\theta}_b}^{\widehat{\theta}_b} \alpha_b \widehat{f}_b^1(x) - (1-\alpha_b)\widehat{f}_b^0(x)dx$$

Combine two cases, let $\Delta f_b^1 = \max_{x\in[\underline{b}^1,\max\{\widetilde{\theta}_b,\widehat{\delta}_b\}]} |\widehat{f}_b^1(x) - \widetilde{f}_b^1(x)|$, $\widehat{L}_b(\widehat{\theta}_b) > \widetilde{L}_b(\widetilde{\theta}_b)$ holds only if the following condition is satisfied:

$$\Delta f_b^1 \alpha_b (\max\{\widetilde{\theta}_b, \widehat{\delta}_b\} - \underline{b}^1) < \int_{\max\{\widetilde{\theta}_b,\widehat{\delta}_b\}}^{\widehat{\theta}_b} \alpha_b \widehat{f}_b^1(x) - (1-\alpha_b)\widehat{f}_b^0(x)dx$$

APPENDIX D

Long-Term Impact of Fairness Interventions on Group Qualification

D.1 Derivations

Qualification profile of a group.

$$\begin{aligned}\gamma_{s,t}(x) &= P_{Y_t|X_t,S}(1|x,s) = \frac{1}{\frac{P_{X_t,Y_t,S}(x,0,s)}{P_{X_t,Y_t,S}(x,1,s)}+1} = \frac{1}{\frac{P_{X_t|Y_t,S}(x|0,s)P_{Y_t|S}(0|s)}{P_{X_t|Y_t,S}(x|1,s)P_{Y_t|S}(1|s)}+1} \\ &= \frac{1}{\frac{P_{X_t|Y_t,S}(x|0,s)}{P_{X_t|Y_t,S}(x|1,s)}(\frac{1}{P_{Y_t|S}(1|s)}-1)+1} = \frac{1}{\frac{f_s^0(x)}{f_s^1(x)}(\frac{1}{\alpha_s(t)}-1)+1}.\end{aligned}$$

Utility of an institute.

$$\mathcal{U}(D_t,Y_t) = \mathbb{E}[R_t(D_t,Y_t)] = n_a\mathbb{E}[R_t(D_t,Y_t)|S=a] + n_b\mathbb{E}[R_t(D_t,Y_t)|S=b]$$

Under policy π^s, we have

$$\begin{aligned}
&\mathbb{E}[R_t(D_t, Y_t)|S = s] = P_{D_t,Y_t|S}(1,1|s)u_+ - P_{D_t,Y_t|S}(1,0|S = s)u_- \\
= &\int_x \Big(P_{D_t,Y_t,X_t|S}(1,1,x|s)u_+ - P_{D_t,Y_t,X_t|S}(1,0,x|s)u_-\Big)dx \\
= &\int_x P_{X_t|S}(x|s)\Big(P_{D_t|X_t,S}(1|x,s)P_{Y_t|X_t,S}(1|x,s)u_+ - P_{D_t|X_t|S}(1|x,s)P_{Y_t|X_t,S}(0|x,s)u_-\Big)dx \\
= &\int_x P_{X_t|S}(x|s)\Big(\pi_s(x)\gamma_{s,t}(x)u_+ - \pi_s(x)(1-\gamma_{s,t}(x))u_-\Big)dx \\
= &\mathbb{E}_{X_t|S=s}[\pi_s(X_t)(\gamma_{s,t}(X_t)(u_+ + u_-) - u_-)].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathcal{U}(D_t, Y_t) = &\; n_a\mathbb{E}_{X_t|S=a}[\pi_a(X_t)(\gamma_{a,t}(X_t)(u_+ + u_-) - u_-)] \\
&+ n_b\mathbb{E}_{X_t|S=b}[\pi_b(X_t)(\gamma_{b,t}(X_t)(u_+ + u_-) - u_-)]
\end{aligned}$$

Dynamics of qualification rate.

$$\begin{aligned}
\alpha_s(t+1) = &\; P_{Y_{t+1}|S}(1|s) = \int_x \sum_{y,a} P_{Y_{t+1},Y_t,D_t,X_t|S}(1,y,d,x|s)dx \\
= &\int_x \sum_{y,a} P_{Y_{t+1}|Y_t,D_t,X_t,S}(1|y,d,x,s)P_{D_t|X_t,S}(d|x,s)P_{X_t|Y_t,S}(x|y,s)P_{Y_t|S}(y|s)dx \\
= &\int_x \sum_{a} \Big\{P_{Y_{t+1}|Y_t,D_t,X_t,S}(1|0,d,x,s)P_{D_t|X_t,S}(d|x,s)P_{X_t|Y_t,S}(x|0,s)\Big\}P_{Y_t|S}(0|s)dx \\
&+ \int_x \sum_{d} \Big\{P_{Y_{t+1}|Y_t,X_t,D_t,S}(1|1,x,d,s)P_{D_t|X_t,S}(d|x,s)P_{X_t|Y_t,S}(x|1,s)\Big\}P_{Y_t|S}(1|s)dx \\
= &\; \mathbb{E}_{X_t|Y_t=0,S=s}\Big[(1-\pi_{s,t}(X_t))T_s^{00} + \pi_{s,t}(X_t)T_s^{01}\Big](1-\alpha_s(t)) \\
&+ \mathbb{E}_{X_t|Y_t=1,S=s}\Big[(1-\pi_{s,t}(X_t))T_{10}^s + \pi_{s,t}(X_t)T_{11}^s\Big]\alpha_s(t) \\
= &\; g_s^0(\alpha_a(t),\alpha_b(t))\cdot(1-\alpha_s(t)) + g_s^1(\alpha_a(t),\alpha_b(t))\cdot\alpha_s(t)
\end{aligned}$$

D.2 Proof that the threshold policies are optimal

In the following proof, we focus on optimal policy at t and omit the subscript t.

First consider unconstrained optimal policy, noted as π_s^{UN}, we have,

$$\pi_s^{\text{UN}} = \arg\max_{\pi_s} \mathbb{E}_{X|S=s}[\pi_s(X)(\gamma_s(X)(u_+ + u_-) - u_-)]$$

Therefore, the optimal policy satisfies $\pi_s^{\text{UN}}(x) = \mathbf{1}(\gamma_s(x) \geq \frac{u_-}{u_+ + u_-})$. Since $\gamma_s(x)$ is monotonically increasing in x under Assumption 9, $\pi_s^{\text{UN}}(x) = \mathbf{1}(x \geq (\gamma_s)^{-1}(\frac{u_-}{u_+ + u_-}))$ is threshold policy where $(\gamma_s)^{-1}(\cdot)$ denotes the inverse function of $\gamma_s(\cdot)$.

Now consider optimal fair policy under some fairness constraint $\mathcal{C}$ satisfying Assumption 10. Consider any pair of policies (π_a, π_b) that satisfies fairness constraint $\mathcal{C}$, and define fairness constant $c = \mathbb{E}_{X\sim\mathcal{P}_a^{\mathcal{C}}}[\pi_a(X)] = \mathbb{E}_{X\sim\mathcal{P}_b^{\mathcal{C}}}[\pi_b(X)] \in [0,1]$. To show the optimal fair policy is threshold policy, we will show that there always exists a pair of threshold policies (π_a^d, π_b^d) such that $\mathbb{E}_{X\sim\mathcal{P}_a^{\mathcal{C}}}[\pi_a^d(X)] = \mathbb{E}_{X\sim\mathcal{P}_b^{\mathcal{C}}}[\pi_b^d(X)] = c$, i.e., the fairness constant is the same as (π_a, π_b), and the utility of (π_a^d, π_b^d) is no less than the utility attained under (π_a, π_b).

$\forall s \in \{a, b\}$, let threshold policy π_s^d be defined such that $\pi_s^d(x) = \mathbf{1}(x \geq \theta_s^d)$ and $\mathbb{E}_{X\sim\mathcal{P}_s^{\mathcal{C}}}[\pi_s^d(X)] = c$ are satisfied. Such policy must exist and the threshold is given by $\theta_s^d = (\mathbb{F}_s^{\mathcal{C}})^{-1}(1-c)$, where $\mathbb{F}_s^{\mathcal{C}}(\theta^s) = \int_{-\infty}^{\theta^s} \mathcal{P}_s^{\mathcal{C}}(x)dx$ is CDF of $\mathcal{P}_s^{\mathcal{C}}$ and $(\mathbb{F}_s^{\mathcal{C}})^{-1}(\cdot)$ is the inverse of it.

Let $R_{\pi_s^d}(D, Y)$, $R_{\pi_s}(D, Y)$ denote the utility attained under policies π_s^d, π_s respectively. Next we will show that $\forall s \in \{a, b\}$, $\mathbb{E}[R_{\pi_s^d}(D, Y) \mid S = s] \geq \mathbb{E}[R_{\pi_s}(D, Y) \mid S = s]$ holds, i.e.,

$$\mathbb{E}_{X|S=s}[\pi_s^d(X)(\gamma_s(X)(u_+ + u_-) - u_-)] \geq \mathbb{E}_{X|S=s}[\pi_s(X)(\gamma_s(X)(u_+ + u_-) - u_-)]$$

Since $\pi_s^d(x) = \mathbf{1}(x \geq \theta_s^d)$, we have the followings,

$$\begin{aligned}
\mathbb{E}_{X|S=s}[\pi_s^d(X)(\gamma_s(X)(u_+ + u_-) - u_-)] &= \int_{\theta_s^d}^{\infty} (\gamma_s(x)(u_+ + u_-) - u_-) P_{X|S}(x|s)dx \\
\mathbb{E}_{X|S=s}[\pi_s(X)(\gamma_s(X)(u_+ + u_-) - u_-)] &= \int_{\theta_s^d}^{\infty} (\gamma_s(x)(u_+ + u_-) - u_-) P_{X|S}(x|s)dx \\
&+ \int_{-\infty}^{\theta_s^d} \pi_s(x)(\gamma_s(x)(u_+ + u_-) - u_-) P_{X|S}(x|s)dx \\
&- \int_{\theta_s^d}^{\infty} (1 - \pi_s(x))(\gamma_s(x)(u_+ + u_-) - u_-) P_{X|S}(x|s)dx
\end{aligned}$$

Since $\mathbb{E}_{X\sim\mathcal{P}_s^{\mathcal{C}}}[\pi_s(X)] = c = \mathbb{E}_{X\sim\mathcal{P}_s^{\mathcal{C}}}[\pi_s^d(X)]$, we have

$$\int_{\theta_s^d}^{\infty}(1-\pi_s(x))\mathcal{P}_s^{\mathcal{C}}(x)dx = \int_{-\infty}^{\theta_s^d}\pi_s(x)\mathcal{P}_s^{\mathcal{C}}(x)dx \tag{D.1}$$

Under Assumption 10, $\frac{P_{X|S}(x|s)}{\mathcal{P}_s^{\mathcal{C}}(x)}$ is non-decreasing. Since $\gamma_s(x) = \alpha_s\frac{f_s^1(x)}{P_{X|S}(x|s)}$ is non-decreasing and $1-\gamma_s(x) = (1-\alpha_s)\frac{f_s^0(x)}{P_{X|S}(x|s)}$ is non-increasing, we have $\frac{f_s^1(x)}{P_{X|S}(x|s)}$ is non-decreasing and $\frac{f_s^0(x)}{P_{X|S}(x|s)}$ is non-increasing. Therefore,

$$\begin{aligned}
&(\gamma_s(x)(u_+ + u_-) - u_-)\frac{P_{X|S}(x|s)}{\mathcal{P}_s^{\mathcal{C}}(x)} = \alpha^s\frac{f_s^1(x)}{\mathcal{P}_s^{\mathcal{C}}(x)}u_+ - (1-\alpha^s)\frac{f_s^0(x)}{\mathcal{P}_s^{\mathcal{C}}(x)}u_- \\
= \quad & \alpha^s\frac{f_s^1(x)}{P_{X|S}(x|s)}\frac{P_{X|S}(x|s)}{\mathcal{P}_s^{\mathcal{C}}(x)}u_+ - (1-\alpha^s)\frac{f_s^0(x)}{P_{X|S}(x|s)}\frac{P_{X|S}(x|s)}{\mathcal{P}_s^{\mathcal{C}}(x)}u_-
\end{aligned}$$

is non-decreasing in x. Combine with Eqn. (D.1), we have the followings,

$$\begin{aligned}
& \int_{-\infty}^{\theta_s^d}\pi_s(x)(\gamma_s(x)(u_+ + u_-) - u_-)P_{X|S}(x|s)dx \\
\leq \quad & \int_{-\infty}^{\theta_s^d}\pi_s(x)(\gamma_s(\theta_s^d)(u_+ + u_-) - u_-)\frac{P_{X|S}(\theta_s^d|s)}{\mathcal{P}_s^{\mathcal{C}}(\theta_s^d)}\mathcal{P}_s^{\mathcal{C}}(x)dx \\
= \quad & \int_{\theta_s^d}^{\infty}(1-\pi_s(x))(\gamma_s(\theta_s^d)(u_+ + u_-) - u_-)\frac{P_{X|S}(\theta_s^d|s)}{\mathcal{P}_s^{\mathcal{C}}(\theta_s^d)}\mathcal{P}_s^{\mathcal{C}}(x)dx \\
\leq \quad & \int_{\theta_s^d}^{\infty}(1-\pi_s(x))(\gamma_s(x)(u_+ + u_-) - u_-)\frac{P_{X|S}(x|s)}{\mathcal{P}_s^{\mathcal{C}}(x)}\mathcal{P}_s^{\mathcal{C}}(x)dx \\
= \quad & \int_{\theta_s^d}^{\infty}(1-\pi_s(x))(\gamma_s(x)(u_+ + u_-) - u_-)P_{X|S}(x|s)dx.
\end{aligned}$$

Therefore, the following holds $\forall s \in \{a, b\}$,

$$\mathbb{E}_{X|S=s}[\pi_s^d(X)(\gamma_s(X)(u_+ + u_-) - u_-)] \geq \mathbb{E}_{X|S=s}[\pi_s(X)(\gamma_s(X)(u_+ + u_-) - u_-)].$$

It shows that the utility attained under threshold policy (π_a^d, π_b^d) is no less than the utility of (π_a, π_b), which concludes that the optimal fair policy $(\pi_a^{\mathcal{C}}, \pi_b^{\mathcal{C}})$ must be threshold policies.

Lemma 21 below further shows that the optimal threshold policy $\theta_s(\alpha_a, \alpha_b)$ is continuous and

non-increasing in α_a and α_b.

Lemma 21. *Let* $\Big(\theta_a(\alpha_a,\alpha_b),\theta_b(\alpha_a,\alpha_b)\Big)$ *be a pair of solutions to Eqn.* (5.3) *under* α_a,α_b. $\forall s\in\{a,b\}$, *if* $\frac{f_s^1(x)}{\mathcal{P}_s^{\mathcal{C}}(x)}$ *and* $\frac{f_s^0(x)}{\mathcal{P}_s^{\mathcal{C}}(x)}$ *are continuous everywhere in x, then* $\theta_s(\alpha_a,\alpha_b)$ *is continuous in both* α_a *and* α_b. *Moreover, under Assumption 10,* $\theta_s(\alpha_a,\alpha_b)$ *is non-increasing in* α_a *and* α_b.

Proof. To prove that *a sufficient condition under which* $\theta_s(\alpha_a,\alpha_b)$ *is continuous in* $\alpha_a,\alpha_b\in[0,1]$ *is that* $\frac{f_s^1(x)}{\mathcal{P}_s^{\mathcal{C}}(x)}$ *and* $\frac{f_s^0(x)}{\mathcal{P}_s^{\mathcal{C}}(x)}$ *are continuous everywhere in x*, we define a function $f_s(\theta_s,\alpha_a,\alpha_b)$:

$$\begin{aligned} f_s(\theta_s,\alpha_a,\alpha_b) &= (\gamma_s(\theta_s)-\frac{u_-}{u_++u_-})\frac{\mathbb{P}(X=\theta_s\mid S=s)}{\mathcal{P}_s^{\mathcal{C}}(\theta_s)} \\ &= [\alpha_s u_+ f_s^1(\theta_s)+\alpha_s u_- f_s^0(\theta_s)-u_- f_s^0(\theta_s)]\frac{1}{\mathcal{P}_s^{\mathcal{C}}(\theta_s)} \\ &= [\alpha_s\frac{f_s^1(\theta_s)}{\mathcal{P}_s^{\mathcal{C}}(\theta_s)}u_+ +(\alpha_s-1)\frac{f_s^0(\theta_s)}{\mathcal{P}_s^{\mathcal{C}}(\theta_s)}u_-]. \end{aligned}$$

According to Equation (5.3), we have $n_a f_a(\theta_a,\alpha_a,\alpha_b)+n_b f_b(\theta_b,\alpha_a,\alpha_b)=0$.

Given any α_a and α_b, and any constant k, let $\tilde{\theta}_s^i$ be one solution to $f_s(\theta_s,\alpha_a,\alpha_b)=k$, where $i=1,...,N$ and N is the number of solutions. Firstly, we show that $\tilde{\theta}_s^i(\alpha_a,\alpha_b)$ is continuous in α_a and α_b, for any $i\in\{1,...,N\}$. Because $\frac{f_s^1(x)}{\mathcal{P}_s^{\mathcal{C}}(x)}$ and $\frac{f_s^0(x)}{\mathcal{P}_s^{\mathcal{C}}(x)}$ are continuous, $f_s(\theta_s,\alpha_a,\alpha_b)$ is continuous in α_a, α_b, and θ_s. Therefore, $\forall\epsilon>0$, $\exists\delta>0$ such that for all $|\alpha_{a'}-\alpha_a|<\delta$ and $|\alpha_{b'}-\alpha_b|<\delta\Longrightarrow|\tilde{\theta}_{s'}^i-\tilde{\theta}_s^i|<\epsilon$. Thus, $\tilde{\theta}_s^i(\alpha_a,\alpha_b)$ is continuous in α_a and α_b, $\forall i\in\{1,...,N\}$.

Next, we show that given α_a and α_b, the solutions to $n_a f_a(\theta_a,\alpha_a,\alpha_b)+n_b f_b(\theta_b,\alpha_a,\alpha_b)=0$ under fairness constraint $\mathcal{C}$ are continuous in α_a and $\alpha_b\in[0,1]$. Under fairness constraints in Equation (5.1), $\theta_a=\eta^{\mathcal{C}}(\theta_b)$ holds for some continuous function $\eta^{\mathcal{C}}(\cdot)$. Consequently, we have $n_a f_a(\eta^{\mathcal{C}}(\theta_b),\alpha_a,\alpha_b)+n_b f_b(\theta_b,\alpha_a,\alpha_b)=0$. Because $f_s(\cdot,\cdot,\cdot)$ and $\eta^{\mathcal{C}}(\cdot)$ are continuous functions, with the same reasoning, given α_a and α_b, the solutions to $n_a f_a(\eta^{\mathcal{C}}(\theta_b),\alpha_a,\alpha_b)+n_b f_b(\theta_b,\alpha_a,\alpha_b)=0$ are continuous in α_a and α_b. In other words, $\theta_s^i(\alpha_a,\alpha_b)$ is continuous.

Under Assumption 10, $f_s(\theta_s,\alpha_a,\alpha_b)$ and $\theta_s(\alpha_a,\alpha_b)$ are continuous. We then prove that if $\frac{f_s^1(x)}{\mathcal{P}_s^{\mathcal{C}}(x)}$ is non-decreasing and $\frac{f_s^0(x)}{\mathcal{P}_s^{\mathcal{C}}(x)}$ is non-increasing in x, then $\theta_s(\alpha_a,\alpha_b)$ is non-increasing in α_a and α_b.

Let $(\eta^{\mathcal{C}}(\theta_b),\theta_b)$ be a pair that satisfies fairness constraint, where $\eta^{\mathcal{C}}(\cdot)$ is some continuous and

strictly increasing function, then the optimal one is the pair that satisfies Equation (5.3) as follows:

$$n_a(\gamma_a(\eta^{\mathcal{C}}(\theta_b)) - \tfrac{u_-}{u_+ + u_-})\frac{P_{X|S}(\eta^{\mathcal{C}}(\theta_b)|a)}{\mathcal{P}_a^{\mathcal{C}}(\eta^{\mathcal{C}}(\theta_b))} + n_b(\gamma_b(\theta_b) - \tfrac{u_-}{u_+ + u_-})\frac{P_{X|S}(\theta_b|b)}{\mathcal{P}_b^{\mathcal{C}}(\theta_b)}$$

$$= n_a\Big[\alpha_a \frac{f_a^1(\eta^{\mathcal{C}}(\theta_b))}{\mathcal{P}_a^{\mathcal{C}}(\eta^{\mathcal{C}}(\theta_b))}u_+ + (\alpha_a - 1)\frac{f_a^0(\eta^{\mathcal{C}}(\theta_b))}{\mathcal{P}_a^{\mathcal{C}}(\eta^{\mathcal{C}}(\theta_b))}u_-\Big] + n_b\Big[\alpha_b \frac{f_b^1(\theta_b)}{\mathcal{P}_b^{\mathcal{C}}(\theta_b)}u_+ + (\alpha_b - 1)\frac{f_b^0(\theta_b)}{\mathcal{P}_b^{\mathcal{C}}(\theta_b)}u_-\Big] = 0.$$

Note that $\forall s \in \{a, b\}$, LHS of above equation is strictly increasing in α_s because the coefficient of α_s is positive. Because $\frac{f_s^1(x)}{\mathcal{P}_s^{\mathcal{C}}(x)}$ is non-decreasing and $\frac{f_s^0(x)}{\mathcal{P}_s^{\mathcal{C}}(x)}$ is non-increasing in x, $\frac{f_s^1(x)}{\mathcal{P}_s^{\mathcal{C}}(x)} - \frac{f_s^0(x)}{\mathcal{P}_s^{\mathcal{C}}(x)}$ is non-decreasing in x. As α_s increases, both $\frac{f_a^1(\eta^{\mathcal{C}}(\theta_b))}{\mathcal{P}_a^{\mathcal{C}}(\eta^{\mathcal{C}}(\theta_b))} - \frac{f_a^0(\eta^{\mathcal{C}}(\theta_b))}{\mathcal{P}_a^{\mathcal{C}}(\eta^{\mathcal{C}}(\theta_b))}$ and $\frac{f_b^1(\theta_b)}{\mathcal{P}_b^{\mathcal{C}}(\theta_b)} - \frac{f_b^0(\theta_b)}{\mathcal{P}_b^{\mathcal{C}}(\theta_b)}$ must not increase so that the optimal fair equation can be maintained. It requires that both θ_b and $\theta_a = \eta^{\mathcal{C}}(\theta_b)$ must not increase. In other words, $\forall s \in \{a, b\}$, $\theta_s(\alpha_a, \alpha_b)$ must be non-increasing in α_a and α_b. □

D.3 Proof of Lemma 7

In the following proof, we focus on optimal policy at t and omit the subscript t.

First consider unconstrained optimal policy. Under threshold policy,

$$\begin{aligned}\theta_s^{\text{UN}} &= \arg\max_{\theta^s} \mathbb{E}_{X|S=s}[\pi_s(X)(\gamma_s(X)(u_+ + u_-) - u_-)] \\ &= \arg\max_{\theta^s} \int_{\theta^s}^{\infty} (\gamma_s(x)(u_+ + u_-) - u_-) P_{X|S}(x|s)dx\end{aligned}$$

Since $\gamma_s(x)$ is monotonically increasing in x under Assumption 9, θ_s^{UN} satisfies $\gamma_s(\theta_s^{\text{UN}}) = \frac{u_-}{u_+ + u_-}$.

Now consider optimal policy under fairness constraint, to satisfy constraint $\mathcal{C}$, $\int_{\theta_a}^{\infty} \mathcal{P}_a^{\mathcal{C}}(x)dx = \int_{\theta_b}^{\infty} \mathcal{P}_b^{\mathcal{C}}(x)dx$ should hold. Denote CDF $\mathbb{F}_s^{\mathcal{C}}(\theta^s) = \int_{-\infty}^{\theta^s} \mathcal{P}_{\mathcal{C}}^s(x)dx$, then for any pair (θ_a, θ_b) that is fair, we have $\theta_a = (\mathbb{F}_a^{\mathcal{C}})^{-1}\mathbb{F}_b^{\mathcal{C}}(\theta_b) = \eta^{\mathcal{C}}(\theta_b)$ hold for some strictly increasing function $\eta^{\mathcal{C}}(\cdot)$. Denote $u = \mathbb{F}_b^{\mathcal{C}}(\theta_b)$ and $\theta_a = (\mathbb{F}_a^{\mathcal{C}})^{-1}(u)$, the following holds,

$$\frac{d\eta^{\mathcal{C}}(\theta_b)}{d\theta_b} = \frac{d(\mathbb{F}_a^{\mathcal{C}})^{-1}\mathbb{F}_b^{\mathcal{C}}(\theta_b)}{d\theta_b} = \frac{d(\mathbb{F}_a^{\mathcal{C}})^{-1}(u)}{du}\frac{du}{d\theta_b} = \frac{1}{(\mathbb{F}_a^{\mathcal{C}})'((\mathbb{F}_a^{\mathcal{C}})^{-1}(u))}\frac{du}{d\theta_b} = \frac{(\mathbb{F}_b^{\mathcal{C}})'(\theta_b)}{(\mathbb{F}_a^{\mathcal{C}})'(\theta_a)} = \frac{\mathcal{P}_b^{\mathcal{C}}(\theta_b)}{\mathcal{P}_a^{\mathcal{C}}(\theta_a)}.$$

Denote $f_s(x) = (\gamma_s(x)(u_+ + u_-) - u_-)P_{X|S}(x|s)$, then we have

$$\theta_b^{\mathcal{C}} = \arg\max_{\theta_b} \mathcal{U}(D, Y) = \arg\max_{\theta_b} \left(n_a \int_{\eta^{\mathcal{C}}(\theta_b)}^{\infty} f^a(x)dx + n_b \int_{\theta_b}^{\infty} f^b(x)dx \right).$$

Let $F(\theta_b) = n_a \int_{\eta^{\mathcal{C}}(\theta_b)}^{\infty} f_a(x)dx + n_b \int_{\theta_b}^{\infty} f_b(x)dx$. Because $\gamma_s(x)$ is monotonically increasing in x under Assumption 9, the optimal $\theta_b^{\mathcal{C}}$ satisfies

$$\begin{aligned}
\frac{dF(\theta_b)}{d\theta_b}\Big|_{\theta_b=\theta_b^{\mathcal{C}}} &= -n_a f^a(\eta^{\mathcal{C}}(\theta_b))\frac{d\eta^{\mathcal{C}}(\theta_b)}{d\theta_b} - n_b f^b(\theta_b)\Big|_{\theta_b=\theta_b^{\mathcal{C}}} \\
&= -n_a(\gamma_a(\eta^{\mathcal{C}}(\theta_b^{\mathcal{C}}))(u_+ + u_-) - u_-)\mathbb{P}(X = \eta^{\mathcal{C}}(\theta_b^{\mathcal{C}}) \mid S = a)\frac{\mathcal{P}_{\mathcal{C}}^b(\theta_b^{\mathcal{C}})}{\mathcal{P}_{\mathcal{C}}^a(\eta^{\mathcal{C}}(\theta_b^{\mathcal{C}}))} \\
&\quad -n_b(\gamma_b(\theta_b^{\mathcal{C}})(u_+ + u_-) - u_-)P_{X|S}(\theta_b^{\mathcal{C}}|b) = 0.
\end{aligned}$$

Therefore,

$$n_a(\gamma_a(\theta_a^{\mathcal{C}})(u_+ + u_-) - u_-)\frac{P_{X|S}(\theta_a^{\mathcal{C}}|a)}{\mathcal{P}_{\mathcal{C}}(\theta_a^{\mathcal{C}})} + n_b(\gamma_b(\theta_b^{\mathcal{C}})(u_+ + u_-) - u_-)\frac{P_{X|S}(\theta_b^{\mathcal{C}}|b)}{\mathcal{P}_{\mathcal{C}}(\theta_b^{\mathcal{C}})} = 0.$$

D.4 Proof of Theorem 17

We define balanced equations and functions for the rest proofs. The dynamics system (5.4) can reach equilibrium if $\alpha_s(t) = \alpha_s(t-1)$ holds. Therefore, the system has equilibrium if there exists solution to the *balanced equations* defined as (D.2).

$$\frac{1}{\alpha_a} - 1 = \frac{1 - g_a^1(\theta^a(\alpha_a, \alpha_b))}{g_a^0(\theta^a(\alpha_a, \alpha_b))}; \quad \frac{1}{\alpha_b} - 1 = \frac{1 - g_b^1(\theta^b(\alpha_a, \alpha_b))}{g_b^0(\theta^b(\alpha_a, \alpha_b))}. \tag{D.2}$$

By removing subscript t and writing threshold θ_s as a function of α_a, α_b, we have $g_s^y(\theta_s(\alpha_a, \alpha_b)) = T_s^{y0}\mathbb{F}_s^y(\theta_s(\alpha_a, \alpha_b)) + T_s^{y1}(1 - \mathbb{F}_s^y(\theta_s(\alpha_a, \alpha_b)))$, denote CDF of $f_s^y(x)$ as $\mathbb{F}_s^y(\theta) = \int_{-\infty}^{\theta} f_s^y(x)dx$.

$\forall s \in \{a, b\}$, let $-s = \{a, b\} \setminus s$. $\forall \alpha_{-s} \in [0, 1]$, define *balanced set* w.r.t. dynamics as $\Psi_s(\alpha_{-s}) = \{\overline{\alpha}_s : \frac{1}{\overline{\alpha}_s} - 1 = \frac{1 - g^{1s}(\theta_s(\overline{\alpha}_s, \alpha_{-s}))}{g^{0s}(\theta_s(\overline{\alpha}_s, \alpha_{-s}))}\}$. If the set size $|\Psi_s(\alpha_{-s})| = 1$ holds $\forall \alpha_{-s} \in [0, 1]$, we define *balanced functions* w.r.t. dynamics as $\Psi_s : [0, 1] \to [0, 1]$ with $\Psi_s(\alpha_{-s}) \in \Psi_s(\alpha_{-s}), \forall \alpha_{-s} \in [0, 1]$.

$\forall s \in \{a,b\}$, define function $l_s(\alpha_s) = \frac{1}{\alpha_s} - 1$ and $h_s(\theta_s(\alpha_a,\alpha_b)) = \frac{1-g_s^1(\theta_s(\alpha_a,\alpha_b))}{g_s^0(\theta_s(\alpha_a,\alpha_b))}$,

$$h_s(\theta_s(\alpha_a,\alpha_b)) = \frac{1-(T_s^{10}\mathbb{F}_s^1(\theta_s(\alpha_a,\alpha_b)) + T_s^{11}(1-\mathbb{F}_s^1(\theta_s(\alpha_a,\alpha_b))))}{T_s^{00}\mathbb{F}_s^0(\theta_s(\alpha_a,\alpha_b)) + T_s^{01}(1-\mathbb{F}_s^0(\theta_s(\alpha_a,\alpha_b)))}.$$

Firstly, we prove that given a fixed $\alpha_{-s} \in [0,1]$ there must exist at least one $\overline{\alpha}_s \in (0,1)$ such that $h_s(\theta_s(\alpha_{-s},\overline{\alpha}_s)) = l^s(\overline{\alpha}_s)$, $s \in \{a,b\}$, $-s = \{a,b\} \setminus s$.

Since $\mathbb{F}_s^y(x)$ is continuous in x, and $\theta_s(\alpha_a,\alpha_b)$ is continuous in α_a and α_b, $\mathbb{F}_s^y(\theta_s(\alpha_a,\alpha_b))$ is continuous in α_a and α_b. Therefore, $h_s(\theta_s(\alpha_a,\alpha_b))$ is continuous in α_a and α_b.

Moreover, $g_s^1(\theta_s(\alpha_a,\alpha_b))$ is the convex combination of T_s^{11} and T_s^{10}, and $g_s^0(\theta_s(\alpha_a,\alpha_b))$ is the convex combination of T_s^{01} and T_s^{00}, the following holds $\forall \alpha_a \in [0,1], \alpha_b \in [0,1]$,

$$\min\{T_s^{10},T_s^{11}\} \le g_s^1(\theta_s(\alpha_a,\alpha_b)) \le \max\{T_s^{10},T_s^{11}\};$$
$$\min\{T_s^{00},T_s^{01}\} \le g_s^0(\theta_s(\alpha_a,\alpha_b)) \le \max\{T_s^{00},T_s^{01}\},$$

which implies $0 < \frac{1-\max\{T_s^{10},T_s^{11}\}}{\max\{T_s^{00},T_s^{01}\}} \le h_s(\theta_s(\alpha_a,\alpha_b)) \le \frac{1-\min\{T_s^{10},T_s^{11}\}}{\min\{T_s^{00},T_s^{01}\}} < +\infty$.

Furthermore, $l_s(\alpha_s) = \frac{1}{\alpha_s} - 1$ is continuous and strictly decreasing in α_s, and

$$\lim_{\alpha_s \to 0} l_s(\alpha_s) = +\infty; \quad \lim_{\alpha_s \to 1} l_s(\alpha_s) = 0,$$

Given a fixed $\alpha_a \in [0,1]$, because $h_b(\theta_b(\alpha_a,\alpha_b))$ is continuous over $\alpha_b \in [0,1]$ and with value varying between $\frac{1-\max\{T_b^{10},T_b^{11}\}}{\max\{T_b^{00},T_b^{01}\}}$ and $\frac{1-\min\{T_b^{10},T_b^{11}\}}{\min\{T_b^{00},T_b^{01}\}}$, and $l_b(\alpha_b)$ is continuous with value varying from $+\infty$ to 0, there must exist at least one $\overline{\alpha}_b \in (0,1)$ such that $h_b(\theta_b(\alpha_a,\overline{\alpha}_b)) = l_b(\overline{\alpha}_b)$. Similarly, given a fixed $\alpha_b \in [0,1]$, there must exist at least one $\overline{\alpha}_a \in (0,1)$ such that $h_a(\theta_a(\overline{\alpha}_a,\alpha_b)) = l_a(\overline{\alpha}_a)$.

Secondly, we prove that all the solutions $(\overline{\alpha}_a,\alpha_b)$ and $(\alpha_a,\overline{\alpha}_b)$ are on continuous curves in the 2D plane $\{(\alpha_a,\alpha_b) : \alpha_a \in [0,1], \alpha_b \in [0,1]\}$.

According to the continuity of $l_s(\cdot)$ and $h_s(\cdot)$, we have $\forall \alpha_a \in [0,1]$, $\lim_{\alpha_{a'} \to \alpha_a} l_a(\alpha_{a'}) = l_a(\alpha_a)$; furthermore, $\forall \alpha_a \in [0,1]$ and $\forall \theta_a^i \in \{\theta_a : l_a(\alpha_a) = h_a(\theta_a)\}$, $\lim_{\theta_{a'}^i \to \theta_a^i} h_a(\theta_{a'}^i) = h_a(\theta_a^i)$. Thus, $\forall \epsilon > 0$, $\exists \delta > 0$, such that $\forall \alpha_a \in [0,1]$, $|\alpha_{a'} - \alpha_a| < \delta \Longrightarrow |\theta_{a'}^i - \theta_a^i| < \epsilon$. Consequently, $\forall \epsilon > 0$, $\exists \delta' > 0$ and $\exists \delta > 0$, such that $\forall \alpha_a \in [0,1]$, $|\alpha_{a'} - \alpha_a| < \delta \Longrightarrow |\theta_{a'}^i - \theta_a^i| < \delta' \Longrightarrow |\alpha_{b'}^i - \alpha_b^i| < \epsilon$, the last statement is because of the continuity of $\theta_a(\alpha_a,\alpha_b)$; in other words, $\forall \alpha_a \in [0,1]$, $\lim_{\alpha_{a'} \to \alpha_a} \alpha_{b'}^i = \alpha_b^i$, where $i = 1,...,N$. Therefore, $(\overline{\alpha}_a,\alpha_b)$ is on a set of continuous curves with α_b varying from 0 to 1. Similarly, one can prove that $(\alpha_a,\overline{\alpha}_b)$ is also on a set of continuous curves with α_a varying from 0 to 1.

Finally, we show the existence of equilibrium $(\widehat{\alpha}_a, \widehat{\alpha}_b)$.

Consider a 2D plane $\{(\alpha_a, \alpha_b) : \alpha_a \in [0,1], \alpha_b \in [0,1]\}$, and $\mathcal{C}_1 = \{(\overline{\alpha}_a, \alpha_b)\}$ and $\mathcal{C}_2 = \{(\alpha_a, \overline{\alpha}_b)\}$ that are two sets of continuous curves in the plane defined earlier. It is straightforward to see that there is at least one curve among $\mathcal{C}_1$ whose α_b varies from 0 to 1 and at least one curve among $\mathcal{C}_2$ whose α_a varies from 0 to 1. These two continuous curves must have at least one intersection. Moreover, this intersection $(\widehat{\alpha}_a, \widehat{\alpha}_b)$ satisfies $h_b(\theta_b(\widehat{\alpha}_a, \widehat{\alpha}_b)) = l_b(\widehat{\alpha}_b)$ and $h_a(\theta_a(\widehat{\alpha}_a, \widehat{\alpha}_b)) = l_a(\widehat{\alpha}_a)$, is an equilibrium of system.

Moreover, we also realized that the proof can also be done by using Brouwer's Fixed Point Theorem in topology.

D.5 Proof of Theorem 18

Following the proof of Theorem 17,

$$h_s(\theta_s(\alpha_a, \alpha_b)) = \frac{1 - g_s^1(\theta_s(\alpha_a, \alpha_b))}{g_s^0(\theta_s(\alpha_a, \alpha_b))} = \frac{1 - (T_s^{10}\mathbb{F}_s^1(\theta_s(\alpha_a, \alpha_b)) + T_s^{11}(1 - \mathbb{F}_s^1(\theta_s(\alpha_a, \alpha_b))))}{T_s^{00}\mathbb{F}_s^0(\theta_s(\alpha_a, \alpha_b)) + T_s^{01}(1 - \mathbb{F}_s^0(\theta_s(\alpha_a, \alpha_b)))}.$$

Note that $\forall y \in \{0,1\}$, $T_s^{y0}\mathbb{F}_s^y(\theta_s(\alpha_a, \alpha_b)) + T_s^{y1}(1 - \mathbb{F}_s^y(\theta_s(\alpha_a, \alpha_b)))$ is the convex combination of T_s^{y0} and T_s^{y1} with CDF $\mathbb{F}_s^y(\theta_s(\alpha_a, \alpha_b))$ as the weight. Because $\mathbb{F}_s^y(\theta_s(\alpha_a, \alpha_b))$ is continuous and non-decreasing in $\theta_s(\alpha_a, \alpha_b)$, under Condition 1a), $h_s(\theta_s(\alpha_a, \alpha_b))$ is non-decreasing in $\theta_s(\alpha_a, \alpha_b)$; while under Condition 1b), $h_s(\theta_s(\alpha_a, \alpha_b))$ is non-increasing in $\theta_s(\alpha_a, \alpha_b)$.

Under unconstrained optimal policy or optimal fair policy with constraint $\mathcal{C}$ satisfying Assumption 9 and 10, $\theta_s(\alpha_a, \alpha_b)$ is non-increasing in α_a, α_b. Therefore, under Condition 1a), $h_s(\theta_s(\alpha_a, \alpha_b))$ is non-decreasing in α_a, α_b, while under Condition 1b), $h_s(\theta_s(\alpha_a, \alpha_b))$ is non-increasing in α_a, α_b. Moreover,

$$\text{Under Condition 1a):} \quad 0 < \frac{1 - T_s^{10}}{T_s^{00}} \le h_s(\theta_s(\alpha_a, \alpha_b)) \le \frac{1 - T_s^{11}}{T_s^{01}} < +\infty$$

$$\text{Under Condition 1b):} \quad 0 < \frac{1 - T_s^{11}}{T_s^{01}} \le h_s(\theta_s(\alpha_a, \alpha_b)) \le \frac{1 - T_s^{10}}{T_s^{00}} < +\infty$$

First consider the case when Condition 1a) is satisfied.

Because function $l_s(\alpha_s) = \frac{1}{\alpha_s} - 1$ is continuous and strictly decreasing from $+\infty$ to 0 over

$\alpha_s \in [0,1]$, $\forall s \in \{a,b\}$. Thus, given any fixed $\alpha_b \in [0,1]$, strictly decreasing function $l_a(\alpha_a)$ and non-decreasing function $h_a(\theta_a(\alpha_a,\alpha_b))$ has exactly one intersection, i.e., $\exists$ only one $\overline{\alpha}_a$ such that $h_a(\theta_a(\overline{\alpha}_a,\alpha_b)) = l_a(\overline{\alpha}_a)$. $\forall \alpha_b$, the set $\Psi_a(\alpha_b) = \{\overline{\alpha}_a : h_a(\theta_a(\overline{\alpha}_a,\alpha_b)) = l_a(\overline{\alpha}_a)\}$ has only one element, and they constitute continuous function $\overline{\alpha}_a = \Psi_a(\alpha_b)$ (balanced function). Similarly, $\forall \alpha_a$, set $\Psi_b(\alpha_a) = \{\overline{\alpha}_b : h_b(\theta_b(\alpha_a,\overline{\alpha}_b)) = l_b(\overline{\alpha}_b)\}$ also has only one element, which forms continuous function $\overline{\alpha}_b = \Psi_b(\alpha_a)$.

Because given any α_a, $h_a(\theta_a(\alpha_a,\alpha_b))$ is non-decreasing in α_b, as α_b increases, the intersection with $l_a(\overline{\alpha}_a)$ is non-increasing. Therefore, $\Psi_a(\alpha_b)$ is non-increasing in α_b. Similarly, $\Psi_b(\alpha_a)$ is also non-increasing in α_a.

On the 2D plane $\{(\alpha_a,\alpha_b) : \alpha_a \in [0,1], \alpha_b \in [0,1]\}$, two curves $\mathcal{C}_1 = \{(\alpha_a,\alpha_b) : \alpha_a = \Psi_a(\alpha_b), \alpha_b \in [0,1]\}$ and $\mathcal{C}_2 = \{(\alpha_a,\alpha_b) : \alpha_b = \Psi_b(\alpha_a), \alpha_a \in [0,1]\}$ are both continuous and non-increasing. One sufficient condition to guarantee $\mathcal{C}_1$ and $\mathcal{C}_2$ have exact one intersection, is that $|\frac{d\Psi_a(\alpha_b)}{d\alpha_b}| < 1, \forall \alpha_b \in [0,1]$ and $|\frac{d\Psi_b(\alpha_a)}{d\alpha_a}| < 1, \forall \alpha_a \in [0,1]$. In the followings, we show these sufficient conditions will hold if $|\frac{\partial h_a(\theta_a(\alpha_a,\alpha_b))}{\partial \alpha_b}| < 1$ and $|\frac{\partial h_b(\theta_b(\alpha_a,\alpha_b))}{\partial \alpha_a}| < 1, \forall \alpha_a, \alpha_b$.

Denote $u = h_a(\theta_a(\Psi_a(\alpha_b),\alpha_b))$, because $l_a(\Psi_a(\alpha_b)) = h_a(\theta_a(\Psi_a(\alpha_b),\alpha_b)), \forall \alpha_b$,

$$\frac{d\Psi_a(\alpha_b)}{d\alpha_b} = \frac{d(l_a)^{-1}(u)}{d\alpha_b} = \frac{d(l_a)^{-1}(u)}{du}\frac{du}{d\alpha_b} = \frac{1}{(l_a)'((l_a)^{-1}(u))}\frac{du}{d\alpha_b} = -((l_a)^{-1}(u))^2\frac{du}{d\alpha_b}.$$

Because $(l_a)^{-1}(u) = \Psi_a(\alpha_b) \in [0,1]$, $-((l_a)^{-1}(u))^2 \in [-1,0]$. Moreover, because of the condition $|\frac{dh_a(\theta_a(\alpha_a,\alpha_b))}{d\alpha_b}| < 1$, we have

$$\left|\frac{d\Psi_a(\alpha_b)}{d\alpha_b}\right| < 1.$$

Similarly, we can show that $|\frac{d\Psi_b(\alpha_a)}{d\alpha_a}| < 1$ holds $\forall \alpha_a$ if $|\frac{\partial h_b(\theta_b(\alpha_a,\alpha_b))}{\partial \alpha_a}| < 1$. Therefore, $\mathcal{C}_1, \mathcal{C}_2$ have only one intersection, the equilibrium $(\widehat{\alpha}_a, \widehat{\alpha}_b)$ is unique.

Now consider the case when Condition 1b) is satisfied.

Because $\frac{dl_s(\alpha_s)}{d\alpha_s} = -\frac{1}{(\alpha_s)^2} < -1, \forall \alpha_s \in (0,1)$, and $-1 \le \frac{\partial h_s(\theta_s(\alpha_a,\alpha_b))}{\partial \alpha_s} \le 0$ for any fixed $\alpha^{-s} \in [0,1]$. Strictly decreasing function $l_s(\alpha_s)$ and non-increasing function $h_s(\theta_s(\alpha_a,\alpha_b))$ has exactly one intersection. Therefore, $\forall \alpha_b$, balanced set $\Psi_a(\alpha_b) = \{\overline{\alpha}_a : h_a(\theta_a(\overline{\alpha}_a,\alpha_b)) = l_a(\overline{\alpha}_a)\}$ has only one element, and they constitute continuous function $\overline{\alpha}_a = \Psi_a(\alpha_b)$ (balanced function). Similarly, $\forall \alpha_b$, set $\Psi_a(\alpha_b) = \{\overline{\alpha}_a : h_a(\theta_a(\overline{\alpha}_a,\alpha_b)) = l_a(\overline{\alpha}_a)\}$ also has only one element, which forms continuous function $\overline{\alpha}_a = \Psi_a(\alpha_b)$.

Because given any α_a, $h_a(\theta_a(\alpha_a,\alpha_b))$ is non-increasing in α_b. As α_b increases, the intersection

with $l_a(\overline{\alpha}_a)$ is non-decreasing. Therefore, $\Psi_a(\alpha_b)$ is non-decreasing in α_b. Similarly, $\Psi_b(\alpha_a)$ is also non-decreasing in α_a.

On the 2D plane $\{(\alpha_a, \alpha_b) : \alpha_a \in [0,1], \alpha_b \in [0,1]\}$, two curves $\mathcal{C}_1 = \{(\alpha_a, \alpha_b) : \alpha_a = \Psi_a(\alpha_b), \alpha_b \in [0,1]\}$ and $\mathcal{C}_2 = \{(\alpha_a, \alpha_b) : \alpha_b = \Psi_b(\alpha_a), \alpha_a \in [0,1]\}$ are both continuous and non-decreasing. One sufficient condition to guarantee $\mathcal{C}_1$ and $\mathcal{C}_2$ have exact one intersection, is that $\frac{d\Psi_a(\alpha_b)}{d\alpha_b} < 1, \forall \alpha_b \in [0,1]$ and $\frac{d\Psi_b(\alpha_a)}{d\alpha_a} < 1, \forall \alpha_a \in [0,1]$. Using the same analysis as the case under Condition 1a), we can show these sufficient conditions will hold if $|\frac{\partial h_a(\theta_a(\alpha_a,\alpha_b))}{\partial \alpha_b}| < 1$ and $|\frac{\partial h_b(\theta_b(\alpha_a,\alpha_b))}{\partial \alpha_a}| < 1, \forall \alpha_a, \alpha_b$.

Therefore, $\mathcal{C}_1, \mathcal{C}_2$ have only one intersection, the equilibrium $(\widehat{\alpha}_a, \widehat{\alpha}_b)$ is unique.

D.6 Proof of Corollary 2.

Define notations $\mathbb{F}_s^y = \mathbb{F}_s^y(\theta_s(\alpha_a, \alpha_b))$, $\Delta T_s^0 = T_s^{01} - T_s^{00}$ and $\Delta T_s^1 = T_s^{11} - T_s^{10}$.

$$h_s(\theta_s(\alpha_a, \alpha_b)) = \frac{(1 - T_s^{10})\mathbb{F}_s^1 + (1 - T_s^{11})(1 - \mathbb{F}_s^1)}{T_s^{00}\mathbb{F}_s^0 + T_s^{01}(1 - \mathbb{F}_s^0)} = \frac{(1 - T_s^{11}) + \Delta T_s^1 \mathbb{F}_s^1}{T_s^{00} + \Delta T_s^0 (1 - \mathbb{F}_s^0)}$$

Take derivative w.r.t. α_u, $\forall u \in \{a, b\}$,

$$\frac{\partial h_s(\theta_s(\alpha_a, \alpha_b))}{\partial \alpha_u} = \frac{\Delta T_s^1 \frac{\partial \mathbb{F}_s^1}{\partial \alpha_u}(T_s^{00} + \Delta T_s^0(1 - \mathbb{F}_s^0)) + \Delta T_s^0 \frac{\partial \mathbb{F}_s^0}{\partial \alpha_u}((1 - T_s^{11}) + \Delta T_s^1 \mathbb{F}_s^1)}{(T_s^{00} + \Delta T_s^0(1 - \mathbb{F}_s^0))^2}$$

Consider case under Condition 1a). Since $\Delta T_s^0 < 0$, $\Delta T_s^1 < 0$, $T_s^{00} + \Delta T_s^0(1 - \mathbb{F}_s^0) > 0$, and $(1 - T_s^{11}) + \Delta T_s^1 \mathbb{F}_s^1 > 0$, we have $|\frac{\partial h_s(\theta_s(\alpha_a,\alpha_b))}{\partial \alpha_u}| \leq |\frac{\Delta T_s^1 M^1 T_s^{00} + \Delta T_s^0 M^0 (1 - T_s^{11})}{(T_s^{01})^2}|$.

Take $\epsilon_s^1 = \epsilon_s^0 = \frac{(T_s^{01})^2}{M^1 T_s^{00} + M^0(1 - T_s^{11})}$, if $|\Delta T_s^1| < \epsilon_s^1$ and $|\Delta T_s^0| < \epsilon_s^0$, then $|\frac{\partial h_s(\theta_s(\alpha_a,\alpha_b))}{\partial \alpha_u}| < 1$ holds. From Theorem 18, the equilibrium of dynamics 5.4 is unique.

Consider case under Condition 1b).

Since $\Delta T_s^0 > 0$ and $\Delta T_s^1 > 0$, we have $|\frac{\partial h_s(\theta_s(\alpha_a,\alpha_b))}{\partial \alpha_u}| \leq \frac{\Delta T_s^1 M^1 T_s^{01} + \Delta T_s^0 M^0 (1 - T_s^{10})}{(T_s^{00})^2}$.

Take $\epsilon_s^1 = \epsilon_s^0 = \frac{(T_s^{00})^2}{M^1 T_s^{01} + M^0(1 - T_s^{10})}$, if $\Delta T_s^1 < \epsilon_s^1$ and $\Delta T_s^0 < \epsilon_s^0$, then $|\frac{\partial h_s(\theta_s(\alpha_a,\alpha_b))}{\partial \alpha_u}| < 1$ holds. From Theorem 18, the equilibrium of dynamics 5.4 is unique.

D.7 The proof of Theorem 19

$\forall s \in \{a,b\}$, an equilibrium $\widehat{\alpha}_s^{\text{UN}}$ satisfies:

$$\frac{1-g_s^1(\theta_s^{\text{UN}}(\widehat{\alpha}_s^{\text{UN}}))}{g_s^0(\theta_s^{\text{UN}}(\widehat{\alpha}_s^{\text{UN}}))} = \frac{1-(T_s^{11}(1-\mathbb{F}_s^1(\theta_s^{\text{UN}}(\widehat{\alpha}_s^{\text{UN}})))+T_s^{10}\mathbb{F}_s^1(\theta_s^{\text{UN}}(\widehat{\alpha}_s^{\text{UN}})))}{T_s^{01}(1-\mathbb{F}_s^0(\theta_s^{\text{UN}}(\widehat{\alpha}_s^{\text{UN}})))+T_s^{00}\mathbb{F}_s^0(\theta_s^{\text{UN}}(\widehat{\alpha}_s^{\text{UN}}))} = \frac{1}{\widehat{\alpha}_s^{\text{UN}}} - 1.$$

One solution to the above equation is:

$$\widehat{\alpha}_s^{\text{UN}} = T_s^{11}(1-\mathbb{F}_s^1(\theta_s^{\text{UN}}(\widehat{\alpha}_s^{\text{UN}})))+T_s^{10}\mathbb{F}_s^1(\theta_s^{\text{UN}}(\widehat{\alpha}_s^{\text{UN}})) = T_s^{01}(1-\mathbb{F}_s^0(\theta_s^{\text{UN}}(\widehat{\alpha}_s^{\text{UN}})))+T_s^{00}\mathbb{F}_s^0(\theta_s^{\text{UN}}(\widehat{\alpha}_s^{\text{UN}}))$$

It shows that $\widehat{\alpha}_s^{\text{UN}}$ is a convex combination of T_s^{00}, T_s^{01}, and also a convex combination of T_s^{10}, T_s^{11}.

$\forall \alpha^{\text{UN}}$ and $\mathbb{F}_s^0(x)$, $\mathbb{F}_s^1(x)$, there is a set of transitions with $T_s^{00} < \alpha^{\text{UN}} < T_s^{01}$ and $T_s^{10} < \alpha^{\text{UN}} < T_s^{11}$ (satisfy Condition 1b)), or $T_s^{01} < \alpha^{\text{UN}} < T_s^{00}$ and $T_s^{11} < \alpha^{\text{UN}} < T_s^{10}$ (satisfy Condition 1a)), such that the above equation holds with $\widehat{\alpha}_s^{\text{UN}} = \alpha^{\text{UN}}$, $\forall s \in \{a,b\}$, i.e., equitable equilibrium is attained.

Next we show that if $f_a^y(x) \neq f_b^y(x)$, then $\widehat{\alpha}_b^{\mathcal{C}} \neq \widehat{\alpha}_a^{\mathcal{C}}$ under these sets of transitions. Under the conditions of Theorem 18, $(\widehat{\alpha}_a^{\mathcal{C}}, \widehat{\alpha}_b^{\mathcal{C}})$ is the intersection of two curves $\mathcal{C}_1 = \{(\alpha_a,\alpha_b) : \alpha_a = \psi_a^{\mathcal{C}}(\alpha_b), \alpha_b \in [0,1]\}$ and $\mathcal{C}_2 = \{(\alpha_a,\alpha_b) : \alpha_b = \psi_b^{\mathcal{C}}(\alpha_a), \alpha_a \in [0,1]\}$; furthermore, let $\widetilde{\alpha}_a^{\mathcal{C}}$, $\widetilde{\alpha}_b^{\mathcal{C}}$ be defined such that $\widetilde{\alpha}_a^{\mathcal{C}} = \psi_a^{\mathcal{C}}(\widetilde{\alpha}_a^{\mathcal{C}})$, $\widetilde{\alpha}_b^{\mathcal{C}} = \psi_b^{\mathcal{C}}(\widetilde{\alpha}_b^{\mathcal{C}})$, which are the intersections of $\alpha_a = \psi_a^{\mathcal{C}}(\alpha_b)$ and $\alpha_a = \alpha_b$, as well as $\alpha_b = \psi_b^{\mathcal{C}}(\alpha_a)$ and $\alpha_a = \alpha_b$, respectively. Then in order to prove $\widehat{\alpha}_b^{\mathcal{C}} \neq \widehat{\alpha}_a^{\mathcal{C}}$, it is sufficient to show $\widetilde{\alpha}_a^{\mathcal{C}} \neq \widetilde{\alpha}_b^{\mathcal{C}}$.

Given $\alpha_a = \alpha_b = \alpha^{\text{UN}}$, because $f_a^y(x) \neq f_b^y(x)$, we have $\theta_s^{\text{UN}}(\alpha^{\text{UN}}) \neq \theta_s^{\mathcal{C}}(\alpha^{\text{UN}},\alpha^{\text{UN}})$ and to satisfy Eqn. (5.3), there are only two possibilities: (1) $\theta_a^{\text{UN}}(\alpha^{\text{UN}}) > \theta_a^{\mathcal{C}}(\alpha^{\text{UN}},\alpha^{\text{UN}})$, $\theta_b^{\text{UN}}(\alpha^{\text{UN}}) < \theta_b^{\mathcal{C}}(\alpha^{\text{UN}},\alpha^{\text{UN}})$; (2) $\theta_a^{\text{UN}}(\alpha^{\text{UN}}) < \theta_a^{\mathcal{C}}(\alpha^{\text{UN}},\alpha^{\text{UN}})$, $\theta_b^{\text{UN}}(\alpha^{\text{UN}}) > \theta_b^{\mathcal{C}}(\alpha^{\text{UN}},\alpha^{\text{UN}})$.

WLOG, suppose the first case holds. Under Condition 1b),

$$\frac{1-g_b^1(\theta_b^{\text{UN}}(\alpha^{\text{UN}}))}{g_b^0(\theta_b^{\text{UN}}(\alpha^{\text{UN}}))} < \frac{1-g_b^1(\theta_b^{\mathcal{C}}(\alpha^{\text{UN}},\alpha^{\text{UN}}))}{g_b^0(\theta_b^{\mathcal{C}}(\alpha^{\text{UN}},\alpha^{\text{UN}}))};\ \ \frac{1-g_a^1(\theta_a^{\text{UN}}(\alpha^{\text{UN}}))}{g_a^0(\theta_a^{\text{UN}}(\alpha^{\text{UN}}))} > \frac{1-g_a^1(\theta_a^{\mathcal{C}}(\alpha^{\text{UN}},\alpha^{\text{UN}}))}{g_a^0(\theta_a^{\mathcal{C}}(\alpha^{\text{UN}},\alpha^{\text{UN}}))}$$

It implies that $\widetilde{\alpha}_b^{\mathcal{C}} < \widehat{\alpha}_b^{\text{UN}} = \widehat{\alpha}_a^{\text{UN}} < \widetilde{\alpha}_a^{\mathcal{C}}$. Similarly, under Condition 1a), $\widetilde{\alpha}_b^{\mathcal{C}} > \widehat{\alpha}_b^{\text{UN}} = \widehat{\alpha}_a^{\text{UN}} > \widetilde{\alpha}_a^{\mathcal{C}}$. Therefore, $\widehat{\alpha}_a^{\mathcal{C}} \neq \widehat{\alpha}_b^{\mathcal{C}}$.

In contrast, if $f_a^y(x) = f_b^y(x)$, we have $\theta_s^{\text{UN}}(\alpha) = \theta_s^{\mathcal{C}}(\alpha,\alpha)$ and $\widetilde{\alpha}_b^{\mathcal{C}} = \widetilde{\alpha}_a^{\mathcal{C}}$. Therefore, $\widehat{\alpha}_a^{\mathcal{C}} = \widehat{\alpha}_b^{\mathcal{C}}$.

D.8 Proof of Theorem 20

WLOG, suppose that $\widehat{\alpha}_a^{\text{UN}} > \widehat{\alpha}_b^{\text{UN}}$ in the proof. Let $\psi_a^{\mathcal{C}}(\cdot), \psi_b^{\mathcal{C}}(\cdot)$ be balanced functions as defined in Theorem 18 under constraint $\mathcal{C}$. Firstly, we show that $\widehat{\alpha}_b^{\text{UN}}$ and $\widehat{\alpha}_a^{\text{UN}}$ are solutions to $\begin{cases} \alpha_b = \psi_b^{\mathcal{C}}(\alpha_a) \\ \alpha_a = \alpha_b \end{cases}$ and $\begin{cases} \alpha_a = \psi_a^{\mathcal{C}}(\alpha_b) \\ \alpha_a = \alpha_b \end{cases}$, respectively, i.e., $\widehat{\alpha}_b^{\text{UN}} = \psi_b^{\mathcal{C}}(\widehat{\alpha}_b^{\text{UN}})$ and $\widehat{\alpha}_a^{\text{UN}} = \psi_a^{\mathcal{C}}(\widehat{\alpha}_a^{\text{UN}})$.

Because $f_a^y(x) = f_b^y(x), \forall y \in \{0,1\}, \forall x$, when $\alpha_a = \alpha_b = \alpha$, we have $\gamma_a(x) = \gamma_b(x)$, $\mathcal{P}_a^{\texttt{EqOpt}}(x) = \mathcal{P}_b^{\texttt{EqOpt}}(x)$ and $\mathcal{P}_a^{\texttt{DP}}(x) = \mathcal{P}_b^{\texttt{DP}}(x)$, which implies $\theta_a^{\mathcal{C}}(\alpha,\alpha) = \theta_b^{\mathcal{C}}(\alpha,\alpha)$; furthermore, the optimal fair policies of `DP` and `EqOpt` satisfy $\gamma_a(\theta_a^{\mathcal{C}}(\alpha,\alpha)) = \gamma_b(\theta_b^{\mathcal{C}}(\alpha,\alpha)) = \frac{u_-}{u_+ + u_-}$ according to the optimal fair policy equation:

$$\frac{n_a\alpha_a}{\gamma_a(\theta_a^{\texttt{EqOpt}})} + \frac{n_b\alpha_b}{\gamma_b(\theta_b^{\texttt{EqOpt}})} = \frac{n_a\alpha_a}{\frac{u_-}{u_+ + u_-}} + \frac{n_b\alpha_b}{\frac{u_-}{u_+ + u_-}}; \quad n_a\gamma_a(\theta_a^{\texttt{DP}}) + n_b\gamma_b(\theta_b^{\texttt{DP}}) = \frac{u_-}{u_+ + u_-}.$$

Because $\gamma_a(\theta_a^{\text{UN}}(\alpha)) = \gamma_b(\theta_b^{\text{UN}}(\alpha)) = \frac{u_-}{u_+ + u_-}$ we have $\gamma_a(\theta_a^{\text{UN}}(\alpha)) = \gamma_a(\theta_a^{\mathcal{C}}(\alpha,\alpha)) = \gamma_b(\theta_b^{\text{UN}}(\alpha)) = \gamma_b(\theta_b^{\mathcal{C}}(\alpha,\alpha))$ so that $\theta_a^{\mathcal{C}}(\alpha,\alpha) = \theta_a^{\text{UN}}(\alpha) = \theta_b^{\mathcal{C}}(\alpha,\alpha) = \theta_b^{\text{UN}}(\alpha)$ holds under any α. $\forall s \in \{a,b\}$, because $\widehat{\alpha}_s^{\text{UN}}$ is the solution to balanced equation, i.e., $l_s(\widehat{\alpha}_s^{\text{UN}}) = h_s(\theta_s^{\text{UN}}(\widehat{\alpha}_s^{\text{UN}}))$. We have $l_s(\widehat{\alpha}_s^{\text{UN}}) = h_s(\theta_s^{\mathcal{C}}(\widehat{\alpha}_s^{\text{UN}}, \widehat{\alpha}_s^{\text{UN}}))$, which further implies $\widehat{\alpha}_s^{\text{UN}} = \psi_s^{\mathcal{C}}(\widehat{\alpha}_s^{\text{UN}})$.

Under Condition 1b), according to the proof of Theorem 18, we know that $0 \le \frac{d\psi_b^{\mathcal{C}}(\alpha_a)}{d\alpha_a} < 1$ and $0 \le \frac{d\psi_a^{\mathcal{C}}(\alpha_b)}{d\alpha_b} < 1$. Because $\widehat{\alpha}_b^{\text{UN}} = \psi_b^{\mathcal{C}}(\widehat{\alpha}_b^{\text{UN}}) < \widehat{\alpha}_a^{\text{UN}} = \psi_a^{\mathcal{C}}(\widehat{\alpha}_a^{\text{UN}})$, we have $\widehat{\alpha}_b^{\text{UN}} < \psi_b^{\mathcal{C}}(\alpha_a) < \alpha_a, \forall \alpha_a \in [\widehat{\alpha}_b^{\text{UN}}, \widehat{\alpha}_a^{\text{UN}}]$. Similarly, we have $\alpha_b < \psi_a^{\mathcal{C}}(\alpha_b) < \widehat{\alpha}_a^{\text{UN}}, \forall \alpha_b \in [\widehat{\alpha}_b^{\text{UN}}, \widehat{\alpha}_a^{\text{UN}}]$. Therefore, after representing the two balanced functions as two curves $\mathcal{C}_1 = \{(\alpha_a, \alpha_b) : \alpha_a = \psi_a^{\mathcal{C}}(\alpha_b), \alpha_b \in [0,1]\}$ and $\mathcal{C}_2 = \{(\alpha_a, \alpha_b) : \alpha_b = \psi_b^{\mathcal{C}}(\alpha_a), \alpha_a \in [0,1]\}$ on the 2D plane $\{(\alpha_a, \alpha_b) : \alpha_a \in [0,1], \alpha_b \in [0,1]\}$, the intersection $(\widehat{\alpha}_a^{\mathcal{C}}, \widehat{\alpha}_b^{\mathcal{C}})$ of $\mathcal{C}_1$ and $\mathcal{C}_2$ satisfies: 1) $\widehat{\alpha}_a^{\mathcal{C}} > \widehat{\alpha}_b^{\mathcal{C}}$; 2) $\widehat{\alpha}_b^{\text{UN}} < \widehat{\alpha}_a^{\mathcal{C}} < \widehat{\alpha}_a^{\text{UN}}$; 3) $\widehat{\alpha}_b^{\text{UN}} < \widehat{\alpha}_b^{\mathcal{C}} < \widehat{\alpha}_a^{\text{UN}}$. Therefore, $|\widehat{\alpha}_a^{\mathcal{C}} - \widehat{\alpha}_b^{\mathcal{C}}| \le |\widehat{\alpha}_a^{\text{UN}} - \widehat{\alpha}_b^{\text{UN}}|$.

Under Condition 1a), according to the proof of Theorem 18, we know that $-1 < \frac{d\psi_b^{\mathcal{C}}(\alpha_a)}{d\alpha_a} \le 0$ and $-1 < \frac{d\psi_a^{\mathcal{C}}(\alpha_b)}{d\alpha_b} \le 0$. Because $\widehat{\alpha}_b^{\text{UN}} = \psi_b^{\mathcal{C}}(\widehat{\alpha}_b^{\text{UN}}) < \widehat{\alpha}_a^{\text{UN}} = \psi_a^{\mathcal{C}}(\widehat{\alpha}_a^{\text{UN}})$, we have $\psi_b^{\mathcal{C}}(\alpha_a) < \widehat{\alpha}_b^{\text{UN}}, \forall \alpha_a > \widehat{\alpha}_b^{\text{UN}}$. Similarly, we have $\psi_a^{\mathcal{C}}(\alpha_b) > \widehat{\alpha}_a^{\text{UN}}, \forall \alpha_b < \widehat{\alpha}_a^{\text{UN}}$. Due to the existence of equilibrium, the intersection $(\widehat{\alpha}_a^{\mathcal{C}}, \widehat{\alpha}_b^{\mathcal{C}})$ of $\mathcal{C}_1$ and $\mathcal{C}_2$ must satisfy: 1) $\widehat{\alpha}_a^{\mathcal{C}} > \widehat{\alpha}_b^{\mathcal{C}}$; 2) $\widehat{\alpha}_a^{\text{UN}} < \widehat{\alpha}_a^{\mathcal{C}}$; 3) $\widehat{\alpha}_b^{\mathcal{C}} < \widehat{\alpha}_b^{\text{UN}}$. Therefore, $|\widehat{\alpha}_a^{\mathcal{C}} - \widehat{\alpha}_b^{\mathcal{C}}| \ge |\widehat{\alpha}_a^{\text{UN}} - \widehat{\alpha}_b^{\text{UN}}|$.

D.9 Proof of Theorem 21

The proof is under the conditions of Theorem 18 such that there is unique equilibrium of qualification rate. Under fairness constraint $\mathcal{C} = \texttt{EqOpt}$ or $\texttt{DP}$, consider 2D plane $\{(\alpha_a, \alpha_b) : \alpha_a \in [0,1], \alpha_b \in [0,1]\}$, and note that equilibrium $(\widetilde{\alpha}_a^{\mathcal{C}}, \widetilde{\alpha}_b^{\mathcal{C}})$ is the intersection of two curves $\mathcal{C}_1 = \{(\alpha_a, \alpha_b) : \alpha_a = \psi_a^{\mathcal{C}}(\alpha_b), \alpha_b \in [0,1]\}$ and $\mathcal{C}_2 = \{(\alpha_a, \alpha_b) : \alpha_b = \psi_b^{\mathcal{C}}(\alpha_a), \alpha_a \in [0,1]\}$. Consider a line $\{(\alpha_a, \alpha_b) : \alpha_a = \alpha_b, \alpha_a \in [0,1], \alpha_b \in [0,1]\}$, which has unique intersection $\widetilde{\alpha}_a^{\mathcal{C}}$ with $\mathcal{C}_1$, and unique intersection $\widetilde{\alpha}_b^{\mathcal{C}}$ with $\mathcal{C}_2$. That is, $\widetilde{\alpha}_a^{\mathcal{C}} = \psi_a^{\mathcal{C}}(\widetilde{\alpha}_a^{\mathcal{C}})$, $\widetilde{\alpha}_b^{\mathcal{C}} = \psi_b^{\mathcal{C}}(\widetilde{\alpha}_b^{\mathcal{C}})$.

First of all, we show that *if* $\frac{u_+}{u_-} \geq \frac{1-T^{10}}{T^{00}}\beta(\widehat{x})$*, under Condition 1b),* $\widehat{\alpha}_b^{UN} < \widehat{\alpha}_a^{UN}$.

By Condition 2, given any $\alpha_a = \alpha_b = \alpha$, the corresponding qualification profiles of $\mathcal{G}_a, \mathcal{G}_b$ satisfy the followings: $\gamma_b(\widehat{x}) = \gamma_a(\widehat{x})$; $\gamma_b(x) < \gamma_a(x), \forall x < \widehat{x}$; $\gamma_b(x) > \gamma_a(x), \forall x > \widehat{x}$. Let $\overline{\alpha}$ be qualification rate such that $\gamma_a(\widehat{x}) = \gamma_b(\widehat{x}) = \frac{u_-}{u_+ + u_-} \Longrightarrow \frac{u_+}{u_-} = \beta(\widehat{x})(\frac{1}{\overline{\alpha}} - 1)$, where $\beta(\widehat{x}) = \frac{f_a^0(\widehat{x})}{f_a^1(\widehat{x})} = \frac{f_b^0(\widehat{x})}{f_b^1(\widehat{x})}$, then $\forall \alpha \in [\overline{\alpha}, 1]$, $\gamma_a(\theta_a^{\text{UN}}(\alpha)) = \gamma_b(\theta_b^{\text{UN}}(\alpha)) = \frac{u_-}{u_+ + u_-} < \frac{1}{\beta(\widehat{x})(\frac{1}{\alpha}-1)+1} = \gamma_a(\widehat{x}) = \gamma_b(\widehat{x})$. Thus, $\forall \alpha \in [\overline{\alpha}, 1]$, $\theta_a^{\text{UN}}(\alpha) < \theta_b^{\text{UN}}(\alpha) < \widehat{x}$, which implies $\mathbb{F}_a^1(\theta_a^{\text{UN}}(\alpha)) < \mathbb{F}_b^1(\theta_b^{\text{UN}}(\alpha))$ and $\mathbb{F}_a^0(\theta_a^{\text{UN}}(\alpha)) < \mathbb{F}_b^0(\theta_b^{\text{UN}}(\alpha))$; furthermore, under Condition 1b), we have

$$\frac{1-T^{11}}{T^{01}} < \frac{1-g_a^1(\theta_a^{\text{UN}}(\alpha))}{g_a^0(\theta_a^{\text{UN}}(\alpha))} < \frac{1-g_b^1(\theta_b^{\text{UN}}(\alpha))}{g_b^0(\theta_b^{\text{UN}}(\alpha))} < \frac{1-T^{10}}{T^{00}}, \ \forall \alpha \in [\overline{\alpha}, 1].$$

Because $\widehat{\alpha}_a^{\text{UN}}$ and $\widehat{\alpha}_b^{\text{UN}}$ are solutions to balance equations, i.e., $\frac{1}{\widehat{\alpha}_a^{\text{UN}}} - 1 = \frac{1-g_a^1(\theta_a^{\text{UN}}(\widehat{\alpha}_a^{\text{UN}}))}{g_a^0(\theta_a^{\text{UN}}(\widehat{\alpha}_a^{\text{UN}}))}$, $\frac{1}{\widehat{\alpha}_b^{\text{UN}}} - 1 = \frac{1-g_b^1(\theta_b^{\text{UN}}(\widehat{\alpha}_b^{\text{UN}}))}{g_b^0(\theta_b^{\text{UN}}(\widehat{\alpha}_b^{\text{UN}}))}$. If $\overline{\alpha} \leq \widehat{\alpha}_b^{\text{UN}}$, the $\widehat{\alpha}_b^{\text{UN}} < \widehat{\alpha}_a^{\text{UN}}$ must hold under Condition 1b). Next, we show that a sufficient condition of $\overline{\alpha} \leq \widehat{\alpha}_b^{\text{UN}}$ is $\frac{u_+}{u_-} \geq \frac{1-T^{10}}{T^{00}}\beta(\widehat{x})$.

$\frac{u_+}{u_-} \geq \frac{1-T^{10}}{T^{00}}\beta(\widehat{x}) \Longrightarrow \frac{1}{\overline{\alpha}} - 1 \geq \frac{1-T^{10}}{T^{00}}$. Since $\frac{1}{\widehat{\alpha}_b^{\text{UN}}} - 1 < \frac{1-T^{10}}{T^{00}}$, we have $\frac{1}{\widehat{\alpha}_b^{\text{UN}}} - 1 < \frac{1}{\overline{\alpha}} - 1$. Thus, $\overline{\alpha} \leq \widehat{\alpha}_b^{\text{UN}}$. Therefore, if $\frac{u_+}{u_-} \geq \frac{1-T^{10}}{T^{00}}\beta(\widehat{x})$, under Condition 1b), $\widehat{\alpha}_b^{\text{UN}} < \widehat{\alpha}_a^{\text{UN}}$.

Fairness constraint `EqOpt`. Secondly, we show that for $\texttt{EqOpt}$ fair policy, if $\frac{u_+}{u_-} \geq \frac{1-T^{10}}{T^{00}}\beta(\widehat{x})$, under Condition 1b), $\widehat{\alpha}_a^{\text{UN}} - \widehat{\alpha}_b^{\text{UN}} > \widehat{\alpha}_a^{\texttt{EqOpt}} - \widehat{\alpha}_b^{\texttt{EqOpt}} \geq 0$. Because two curves $\mathcal{C}_1, \mathcal{C}_2$ are monotonic increasing. It's sufficient to show two parts: (1) $\widetilde{\alpha}_a^{\texttt{EqOpt}} < \widehat{\alpha}_a^{\text{UN}}$, $\widetilde{\alpha}_b^{\texttt{EqOpt}} > \widehat{\alpha}_b^{\text{UN}}$; (2) $\widetilde{\alpha}_a^{\texttt{EqOpt}} \geq \widetilde{\alpha}_b^{\texttt{EqOpt}}$.

Under $\texttt{EqOpt}$ constraint, $\forall \alpha_a, \alpha_b$, $\mathbb{F}_a^1(\theta_a^{\texttt{EqOpt}}(\alpha_a, \alpha_b)) = \mathbb{F}_b^1(\theta_b^{\texttt{EqOpt}}(\alpha_a, \alpha_b))$ must hold so that $\theta_a^{\texttt{EqOpt}}(\alpha_a, \alpha_b) = \theta_b^{\texttt{EqOpt}}(\alpha_a, \alpha_b)$. Consider the case $\alpha_a = \alpha_b = \alpha$, $\forall \alpha \geq \overline{\alpha}$, we have $\theta_a^{\texttt{EqOpt}}(\alpha, \alpha) = \theta_b^{\texttt{EqOpt}}(\alpha, \alpha)$ and $\theta_a^{\text{UN}}(\alpha) < \theta_b^{\text{UN}}(\alpha)$. It implies that $\theta_a^{\text{UN}}(\alpha) < \theta_a^{\texttt{EqOpt}}(\alpha, \alpha) = \theta_b^{\texttt{EqOpt}}(\alpha, \alpha) < \theta_b^{\text{UN}}(\alpha) < \widehat{x}$,

otherwise Equation (5.3) will be violated. Therefore, the followings hold $\forall \alpha \in [\overline{\alpha}, 1]$,

$$\frac{1-g_a^1(\theta_a^{\texttt{EqOpt}}(\alpha,\alpha))}{g_a^0(\theta_a^{\texttt{EqOpt}}(\alpha,\alpha))} > \frac{1-g_a^1(\theta_a^{\texttt{UN}}(\alpha))}{g_a^0(\theta_a^{\texttt{UN}}(\alpha))}; \quad \frac{1-g_b^1(\theta_b^{\texttt{EqOpt}}(\alpha,\alpha))}{g_b^0(\theta_b^{\texttt{EqOpt}}(\alpha,\alpha))} < \frac{1-g_b^1(\theta_b^{\texttt{UN}}(\alpha))}{g_b^0(\theta_b^{\texttt{UN}}(\alpha))}.$$

$\forall s \in \{a,b\}$, $\widetilde{\alpha}_s^{\texttt{EqOpt}}$ is the solution to $\frac{1-g_s^1(\theta_s^{\texttt{EqOpt}}(\alpha,\alpha))}{g_s^0(\theta_s^{\texttt{EqOpt}}(\alpha,\alpha))} = \frac{1}{\alpha} - 1$ while $\widehat{\alpha}_s^{\texttt{UN}}$ is the solution to $\frac{1-g_s^1(\theta_s^{\texttt{UN}}(\alpha))}{g_s^0(\theta_s^{\texttt{UN}}(\alpha))} = \frac{1}{\alpha} - 1$. Since $\overline{\alpha} \le \widehat{\alpha}_b^{\texttt{UN}} < \widehat{\alpha}_a^{\texttt{UN}}$, it implies $\widetilde{\alpha}_a^{\texttt{EqOpt}} < \widehat{\alpha}_a^{\texttt{UN}}$, $\widetilde{\alpha}_b^{\texttt{EqOpt}} > \widehat{\alpha}_b^{\texttt{UN}}$.

Next, show that $\widetilde{\alpha}_a^{\texttt{EqOpt}} \ge \widetilde{\alpha}_b^{\texttt{EqOpt}}$. $\forall \alpha \ge \overline{\alpha}$, $\theta_a^{\texttt{EqOpt}}(\alpha,\alpha) = \theta_b^{\texttt{EqOpt}}(\alpha,\alpha)$ implies $\mathbb{F}_a^1(\theta_a^{\texttt{EqOpt}}(\alpha,\alpha)) = \mathbb{F}_b^1(\theta_b^{\texttt{EqOpt}}(\alpha,\alpha))$ and $\mathbb{F}_a^0(\theta_a^{\texttt{EqOpt}}(\alpha,\alpha)) \le \mathbb{F}_b^0(\theta_b^{\texttt{EqOpt}}(\alpha,\alpha))$. Therefore,

$$\frac{1-g_a^1(\theta_a^{\texttt{EqOpt}}(\alpha,\alpha))}{g_a^0(\theta_a^{\texttt{EqOpt}}(\alpha,\alpha))} \le \frac{1-g_b^1(\theta_b^{\texttt{EqOpt}}(\alpha,\alpha))}{g_b^0(\theta_b^{\texttt{EqOpt}}(\alpha,\alpha))}.$$

Intersections with function $\frac{1}{\alpha} - 1$ satisfies $\widetilde{\alpha}_a^{\texttt{EqOpt}} \ge \widetilde{\alpha}_b^{\texttt{EqOpt}}$.

It thus concludes that $\widehat{\alpha}_a^{\texttt{UN}} - \widehat{\alpha}_b^{\texttt{UN}} > \widetilde{\alpha}_a^{\texttt{EqOpt}} - \widetilde{\alpha}_b^{\texttt{EqOpt}} \ge 0$.

Fairness constraint `DP`. Finally, consider `DP` fair policy, where $\forall \alpha_a, \alpha_b$,

$(1-\alpha_a)\mathbb{F}_a^0(\theta_a^{\texttt{DP}}(\alpha_a,\alpha_b)) + \alpha_a \mathbb{F}_a^1(\theta_a^{\texttt{DP}}(\alpha_a,\alpha_b)) = (1-\alpha_b)\mathbb{F}_b^0(\theta_b^{\texttt{DP}}(\alpha_a,\alpha_b)) + \alpha_b \mathbb{F}_b^1(\theta_b^{\texttt{DP}}(\alpha_a,\alpha_b))$ must hold.

We first show that under Condition 1b), $\widetilde{\alpha}_a^{\texttt{DP}} < \widehat{\alpha}_a^{\texttt{UN}}$, $\widetilde{\alpha}_b^{\texttt{DP}} > \widehat{\alpha}_b^{\texttt{UN}}$. Consider the case $\alpha_a = \alpha_b = \alpha$, $\forall \alpha \ge \overline{\alpha}$. Since $\forall x$, $(1-\alpha)\mathbb{F}_b^0(x) + \alpha \mathbb{F}_b^1(x) \ge (1-\alpha)\mathbb{F}_a^0(x) + \alpha \mathbb{F}_a^1(x)$, $(1-\alpha)\mathbb{F}_a^0(\theta_a^{\texttt{DP}}(\alpha,\alpha)) + \alpha \mathbb{F}_a^1(\theta_a^{\texttt{DP}}(\alpha,\alpha)) = (1-\alpha)\mathbb{F}_b^0(\theta_b^{\texttt{DP}}(\alpha,\alpha)) + \alpha \mathbb{F}_b^1(\theta_b^{\texttt{DP}}(\alpha,\alpha))$ implies $\theta_a^{\texttt{DP}}(\alpha,\alpha) \ge \theta_b^{\texttt{DP}}(\alpha,\alpha)$. Because $\theta_a^{\texttt{UN}}(\alpha) < \theta_b^{\texttt{UN}}(\alpha)$, $\forall \alpha \ge \overline{\alpha}$. It implies that $\theta_a^{\texttt{DP}}(\alpha,\alpha) > \theta_a^{\texttt{UN}}(\alpha)$ and $\widehat{x} > \theta_b^{\texttt{UN}}(\alpha) > \theta_b^{\texttt{DP}}(\alpha,\alpha)$ must hold. Therefore, $\forall \alpha \in [\overline{\alpha}, 1]$,

$$\frac{1-g_a^1(\theta_a^{\texttt{DP}}(\alpha,\alpha))}{g_a^0(\theta_a^{\texttt{DP}}(\alpha,\alpha))} > \frac{1-g_a^1(\theta_a^{\texttt{UN}}(\alpha))}{g_a^0(\theta_a^{\texttt{UN}}(\alpha))}; \quad \frac{1-g_b^1(\theta_b^{\texttt{DP}}(\alpha,\alpha))}{g_b^0(\theta_b^{\texttt{DP}}(\alpha,\alpha))} < \frac{1-g_b^1(\theta_b^{\texttt{UN}}(\alpha))}{g_b^0(\theta_b^{\texttt{UN}}(\alpha))}$$

Similar to reasoning in `EqOpt` case, we have $\widetilde{\alpha}_a^{\texttt{DP}} < \widehat{\alpha}_a^{\texttt{UN}}$, $\widetilde{\alpha}_b^{\texttt{DP}} > \widehat{\alpha}_b^{\texttt{UN}}$.

Different from `EqOpt` fairness where $\widetilde{\alpha}_a^{\texttt{EqOpt}} \ge \widetilde{\alpha}_b^{\texttt{EqOpt}}$, both $\widetilde{\alpha}_a^{\texttt{DP}} \ge \widetilde{\alpha}_b^{\texttt{DP}}$ and $\widetilde{\alpha}_a^{\texttt{DP}} \le \widetilde{\alpha}_b^{\texttt{DP}}$ are likely to occur, depending on distributions $f_a^0(x)$, $f_b^0(x)$, $f_a^1(x)$ and $f_b^1(x)$. It is because $\theta_a^{\texttt{DP}}(\alpha,\alpha) > \theta_b^{\texttt{DP}}(\alpha,\alpha)$ can result in either $\mathbb{F}_a^0(\theta_a^{\texttt{DP}}(\alpha,\alpha)) \le \mathbb{F}_b^0(\theta_b^{\texttt{DP}}(\alpha,\alpha))$ or $\mathbb{F}_a^0(\theta_a^{\texttt{DP}}(\alpha,\alpha)) \ge \mathbb{F}_b^0(\theta_b^{\texttt{DP}}(\alpha,\alpha))$.

For these two outcomes, if $\widetilde{\alpha}_a^{\texttt{DP}} \ge \widetilde{\alpha}_b^{\texttt{DP}}$, then `DP` fair policy results in a more equitable equilibrium

than unconstrained policy; if $\widetilde{\alpha}_a^{\texttt{DP}} \leq \widetilde{\alpha}_b^{\texttt{DP}}$, it means the disadvantaged group is flipped from $\mathcal{G}_b$ to $\mathcal{G}_a$.

D.10 Proof of Proposition 3

In the proof, we simplify the notations by removing subscript $\mathcal{C}$.

Let $\psi_s(\cdot)$, $\psi_{s'}(\cdot)$ be balanced function of policies (θ_a, θ_b) and $(\theta_{a'}, \theta_{b'})$, respectively.

According to the balanced equation (D.2),

$$\frac{1}{\alpha_s} - 1 = \frac{1 - g_s^1(\theta_s(\alpha_a, \alpha_b))}{g_s^0(\theta_s(\alpha_a, \alpha_b))} = \frac{1 - (T_s^{11}(1 - \mathbb{F}_s^1(\theta_s(\alpha_a, \alpha_b))) + T_s^{10}\mathbb{F}_s^1(\theta_s(\alpha_a, \alpha_b)))}{T_s^{01}(1 - \mathbb{F}_s^0(\theta_s(\alpha_a, \alpha_b))) + T_s^{00}\mathbb{F}_s^0(\theta_s(\alpha_a, \alpha_b))}.$$

Under Condition b), $\forall \alpha_a, \alpha_b \in [0,1]$, $\theta_{a'}(\alpha_a, \alpha_b) < \theta_a(\alpha_a, \alpha_b)$ and $\theta_{b'}(\alpha_a, \alpha_b) < \theta_b(\alpha_a, \alpha_b)$.

Under Condition a), $\forall \alpha_a, \alpha_b \in [0,1]$, $\theta_{a'}(\alpha_a, \alpha_b) > \theta_a(\alpha_a, \alpha_b)$ and $\theta_{b'}(\alpha_a, \alpha_b) > \theta_b(\alpha_a, \alpha_b)$.

Both imply that $\frac{1-g_s^1(\theta_s(\alpha_a,\alpha_b))}{g_s^0(\theta_s(\alpha_a,\alpha_b))} > \frac{1-g_s^1(\theta_{s'}(\alpha_a,\alpha_b))}{g_s^0(\theta_{s'}(\alpha_a,\alpha_b))}$, and $\forall \alpha_a, \alpha_b \in [0,1]$, $\psi_a(\alpha_b) < \psi_{a'}(\alpha_b)$ and $\psi_b(\alpha_a) < \psi_{b'}(\alpha_a)$ hold. As a consequence, $\widehat{\alpha}_{a'} > \widehat{\alpha}_a$ and $\widehat{\alpha}_{b'} > \widehat{\alpha}_b$.

Now consider the long-run average utility of institute $\overline{U}(\theta_a, \theta_b) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathcal{U}_t(\theta_a, \theta_b)$, where the instantaneous utility at t under threshold policies θ_a, θ_b is

$$\begin{aligned}
\mathcal{U}_t(\theta_a, \theta_b) &= \sum_{s=a,b} n_s \mathbb{E}_{X_t|S=s}[\mathbf{1}(X_t \geq \theta_s)(\gamma_{s,t}(X_t)(u_+ + u_-) - u_-)] \\
&= \sum_{s=a,b} n_s \int_{\theta_s}^{\infty} (\gamma_{s,t}(x)(u_+ + u_-) - u_-) P_{X_t|S}(x|s) dx \\
&= \sum_{s=a,b} n_s \int_{\theta_s}^{\infty} \alpha_s(t)(f_s^1(x)u_+ + f_s^0(x)u_-) - f_s^0(x)u_- dx
\end{aligned}$$

In the followings, we use a special case ($\mathcal{C} = \texttt{EqOpt}$, $f_a^y(x) = f_b^y(x), \forall x, y = 0, 1$, under Condition 1b)) to show that $\overline{U}(\theta_{a'}, \theta_{b'}) > \overline{U}(\theta_a, \theta_b)$ can be attained, i.e., the long-run average utility under policy $(\theta_{a'}, \theta_{b'})$ can be higher than myopic optimal policy (θ_a, θ_b).

Since the qualification rates of two groups converge to equilibrium, $\overline{U}(\theta_a, \theta_b) = \mathcal{U}_\infty(\theta_a, \theta_b)$ is the same as instantaneous expected utility of institute at the equilibrium state. To show that $\overline{U}(\theta_{a'}, \theta_{b'}) > \overline{U}(\theta_a, \theta_b)$, we prove the following holds,

$$\sum_{s=a,b} n_s \int_{\theta_{s'}(\widehat{\alpha}_{a'}, \widehat{\alpha}_{b'})}^{\infty} f(x; \widehat{\alpha}_{s'}) dx > \sum_{s=a,b} n_s \int_{\theta_s(\widehat{\alpha}_a, \widehat{\alpha}_b)}^{\infty} f(x; \widehat{\alpha}_s) dx \tag{D.3}$$

where $f(x;\widehat{\alpha}_s) = \widehat{\alpha}_s(f_s^1(x)u_+ + f_s^0(x)u_-) - f_s^0(x)u_-$.

Because $\widehat{\alpha}_{s'} > \widehat{\alpha}_s$, $\theta_{s'}(\widehat{\alpha}_{a'},\widehat{\alpha}_{b'}) < \theta_s(\widehat{\alpha}_{a'},\widehat{\alpha}_{b'}) < \theta_s(\widehat{\alpha}_a,\widehat{\alpha}_b)$ holds under Condition b). LHS of above inequality can be written as

$$\sum_{s=a,b} n_s\Big(\int_{\theta_{s'}(\widehat{\alpha}_{a'},\widehat{\alpha}_{b'})}^{\theta_s(\widehat{\alpha}_a,\widehat{\alpha}_b)} f(x;\widehat{\alpha}_{s'})dx + \int_{\theta_s(\widehat{\alpha}_a,\widehat{\alpha}_b)}^{\infty} f(x;\widehat{\alpha}_{s'})dx\Big).$$

Inequality (D.3) can further be re-organized,

$$\sum_{s=a,b} n_s \int_{\theta_{s'}(\widehat{\alpha}_{a'},\widehat{\alpha}_{b'})}^{\theta_s(\widehat{\alpha}_a,\widehat{\alpha}_b)} f(x;\widehat{\alpha}_{s'})dx > \sum_{s=a,b} n_s \int_{\theta_s(\widehat{\alpha}_a,\widehat{\alpha}_b)}^{\infty} (f(x;\widehat{\alpha}_s) - f(x;\widehat{\alpha}_{s'}))dx \tag{D.4}$$

Consider a special case where $\mathcal{C} = \texttt{EqOpt}$ and $f_a^y(x) = f_b^y(x) = f^y(x), \forall x, \forall y \in \{0,1\}$. Then $\forall \alpha_a, \alpha_b$, we have $\theta_a(\alpha_a,\alpha_b) = \theta_b(\alpha_a,\alpha_b)$ and $\theta_{a'}(\alpha_a,\alpha_b) = \theta_{b'}(\alpha_a,\alpha_b)$. Inequality (D.4) can be reduced to the following, $\forall s \in \{a,b\}$, simplify notations and let $\widehat{\theta} = \theta_s(\widehat{\alpha}_a,\widehat{\alpha}_b)$, $\widehat{\theta'} = \theta_{s'}(\widehat{\alpha}_{a'},\widehat{\alpha}_{b'})$.

$$\begin{aligned}
& \Big(n_a\widehat{\alpha}_{a'} + n_b\widehat{\alpha}_{b'}\Big)\Big(\mathbb{F}^1(\widehat{\theta}) - \mathbb{F}^1(\widehat{\theta'})\Big)u_+ \\
+ \; & \underbrace{\Big(u_+(1-\mathbb{F}^1(\widehat{\theta})) + u_-(1-\mathbb{F}^0(\widehat{\theta}))\Big)\Big(n_a(\widehat{\alpha}_{a'} - \widehat{\alpha}_a) + n_b(\widehat{\alpha}_{b'} - \widehat{\alpha}_b)\Big)}_{\textbf{term 1}} \\
> \; & \Big(n_a(1-\widehat{\alpha}_{a'}) + n_b(1-\widehat{\alpha}_{b'})\Big)\Big(\mathbb{F}^0(\widehat{\theta}) - \mathbb{F}^0(\widehat{\theta'})\Big)u_-
\end{aligned} \tag{D.5}$$

Because $\frac{1}{\widehat{\alpha}_{s'}} - 1 = \frac{1-g_s^1(\widehat{\theta'})}{g_s^0(\widehat{\theta'})}$ and $\frac{1}{\widehat{\alpha}_s} - 1 = \frac{1-g_s^1(\widehat{\theta})}{g_s^0(\widehat{\theta})}$.

$$\widehat{\alpha}_{s'} - \widehat{\alpha}_s > \frac{T_s^{01} - T_s^{00}}{1 - T_s^{10} + T_s^{01}}(\mathbb{F}^0(\widehat{\theta}) - \mathbb{F}^0(\widehat{\theta'}))$$

We have **term 1** >

$$\underbrace{\Big(\frac{u_+}{u_-}(1-\mathbb{F}^1(\widehat{\theta})) + (1-\mathbb{F}^0(\widehat{\theta}))\Big)\Big(n_a\frac{T_a^{01} - T_a^{00}}{1 - T_a^{10} + T_a^{01}} + n_b\frac{T_b^{01} - T_b^{00}}{1 - T_b^{10} + T_b^{01}}\Big)}_{=h(\widehat{\theta})>0}\Big(\mathbb{F}^0(\widehat{\theta}) - \mathbb{F}^0(\widehat{\theta'})\Big)u_-$$

For the optimal `EqOpt` fair threshold $\theta(\widehat{\alpha}_{a'}, \widehat{\alpha}_{b'})$, the following holds

$$\begin{aligned}
\Big(n_a\widehat{\alpha}_{a'} + n_b\widehat{\alpha}_{b'}\Big)f^1(\theta(\widehat{\alpha}_{a'}, \widehat{\alpha}_{b'}))u_+ &= \Big(n_a(1-\widehat{\alpha}_{a'}) + n_b(1-\widehat{\alpha}_{b'})\Big)f^0(\theta(\widehat{\alpha}_{a'}, \widehat{\alpha}_{b'}))u_- \\
\Big(n_a\widehat{\alpha}_{a'} + n_b\widehat{\alpha}_{b'}\Big)f^1(x)u_+ &> \Big(n_a(1-\widehat{\alpha}_{a'}) + n_b(1-\widehat{\alpha}_{b'})\Big)f^0(x)u_-, \forall x > \theta(\widehat{\alpha}_{a'}, \widehat{\alpha}_{b'}) \\
\Big(n_a\widehat{\alpha}_{a'} + n_b\widehat{\alpha}_{b'}\Big)f^1(x)u_+ &< \Big(n_a(1-\widehat{\alpha}_{a'}) + n_b(1-\widehat{\alpha}_{b'})\Big)f^0(x)u_-, \forall x < \theta(\widehat{\alpha}_{a'}, \widehat{\alpha}_{b'})
\end{aligned}$$

It implies that $\exists$ some $\delta > 0$ s.t. $\forall x \in (\theta(\widehat{\alpha}_{a'}, \widehat{\alpha}_{b'}) - \delta, \theta(\widehat{\alpha}_{a'}, \widehat{\alpha}_{b'}) + \delta) = \mathcal{B}(\theta(\widehat{\alpha}_{a'}, \widehat{\alpha}_{b'}), \delta)$,

$$\Big(n_a\widehat{\alpha}_{a'} + n_b\widehat{\alpha}_{b'}\Big)f^1(x)u_+ + h(\widehat{\theta})f^0(x)u_- > \Big(n_a(1-\widehat{\alpha}_{a'}) + n_b(1-\widehat{\alpha}_{b'})\Big)f^0(x)u_-.$$

$\widehat{\theta}, \widehat{\theta'} \in \mathcal{B}(\theta(\widehat{\alpha}_{a'}, \widehat{\alpha}_{b'}), \delta)$ can be satisfied as long as $|\theta_s(\alpha_a, \alpha_b) - \theta_{s'}(\alpha_a, \alpha_b)| \leq \epsilon$ for some sufficiently small $\epsilon > 0$.

Using the mean value theorem, $\exists f^y(x)$ and $\widetilde{x} \in (\widehat{\theta'}, \widehat{\theta}) \subset \mathcal{B}(\theta(\widehat{\alpha}_{a'}, \widehat{\alpha}_{b'}), \delta)$ s.t.

$$\begin{aligned}
& \Big(n_a\widehat{\alpha}_{a'} + n_b\widehat{\alpha}_{b'}\Big)(\mathbb{F}^1(\widehat{\theta}) - \mathbb{F}^1(\widehat{\theta'}))u_+ + h(\widehat{\theta})(\mathbb{F}^0(\widehat{\theta}) - \mathbb{F}^0(\widehat{\theta'}))u_- \\
= & \Big(\Big(n_a\widehat{\alpha}_{a'} + n_b\widehat{\alpha}_{b'}\Big)f^1(\widetilde{x})u_+ + h(\widehat{\theta})f^0(\widetilde{x})u_-\Big)(\widehat{\theta} - \widehat{\theta'}) \\
> & \Big(\Big(n_a(1-\widehat{\alpha}_{a'}) + n_b(1-\widehat{\alpha}_{b'})\Big)f^0(\widetilde{x})u_-\Big)(\widehat{\theta} - \widehat{\theta'}) \\
\geq & \Big(n_a(1-\widehat{\alpha}_{a'}) + n_b(1-\widehat{\alpha}_{b'})\Big)(\mathbb{F}^0(\widehat{\theta}) - \mathbb{F}^0(\widehat{\theta'}))u_-.
\end{aligned}$$

Therefore, inequality (D.5) holds and $\overline{U}(\theta_{a'}, \theta_{b'}) > \overline{U}(\theta_a, \theta_b)$.

D.11 Proof of Proposition 4

To ensure $\alpha_s(t) \to \widehat{\alpha}$, threshold policy $\theta_s(\alpha_s)$ as a function of $\alpha_s \in [0,1]$ should be designed such that $\frac{1-g_s^1(\theta_s(\alpha_s))}{g_s^0(\theta_s(\alpha_s))} = \frac{1}{\alpha_s} - 1$ has a unique solution $\widehat{\alpha}$. Let $\mathcal{I}_s = [\frac{1-\max\{T_s^{11}, T_s^{10}\}}{\max\{T_s^{01}, T_s^{00}\}}, \frac{1-\min\{T_s^{11}, T_s^{10}\}}{\min\{T_s^{01}, T_s^{00}\}}]$, then $\frac{1-g_s^1(\theta_s(\alpha_s))}{g_s^0(\theta_s(\alpha_s))} \in \mathcal{I}_s$ for any threshold policy $\theta_s(\alpha_s)$.

If $\mathcal{I}_a \cap \mathcal{I}_b = \emptyset$, then $\frac{1-g_a^1(\theta_a(\alpha))}{g_a^0(\theta_a(\alpha))} = \frac{1-g_b^1(\theta_b(\alpha))}{g_b^0(\theta_b(\alpha))}$ can never be attained, i.e., no threshold policy can result in equitable equilibrium.

If $\mathcal{I}_a \cap \mathcal{I}_b \neq \emptyset$, then $\forall \widehat{\alpha} \in \mathcal{I}_a \cap \mathcal{I}_b$ and $\forall s \in \{a, b\}$, there exists threshold policy $\theta_s(\alpha_s)$ such that

$\frac{1-g_s^1(\theta_s(\widehat{\alpha}))}{g_s^0(\theta_s(\widehat{\alpha}))} = \frac{1}{\widehat{\alpha}} - 1$. Specifically, under Condition 1b) (resp. 1a)), function

$$h_s(x) = \frac{1-g_s^1(x)}{g_s^0(x)} = \frac{1-(T_s^{11}(1-\mathbb{F}_s^1(x)) + T_s^{10}\mathbb{F}_s^1(x))}{T_s^{01}(1-\mathbb{F}_s^0(x)) + T_s^{00}\mathbb{F}_s^0(x)}$$

is strictly increasing (resp. decreasing) in $x \in (-\infty, +\infty)$ from $\frac{1-T_s^{11}}{T_s^{01}}$ (resp. $\frac{1-T_s^{10}}{T_s^{00}}$) to $\frac{1-T_s^{10}}{T_s^{00}}$ (resp. $\frac{1-T_s^{11}}{T_s^{01}}$) and any non-increasing function $\theta_s(\alpha_s)$ that satisfies $\theta_s(\widehat{\alpha}) = (h_s)^{-1}(\frac{1}{\widehat{\alpha}} - 1)$ can result in $\alpha_s(t) \to \widehat{\alpha}$, where $(h_s)^{-1}(\cdot)$ is the inverse function of $h_s(\cdot)$.

D.12 Proof of Proposition 5

According to the balanced equation (D.2),

$$\frac{1}{\alpha_s} - 1 = \frac{1-g_s^1(\theta_s(\alpha_a,\alpha_b))}{g_s^0(\theta_s(\alpha_a,\alpha_b))} = \frac{1-(T_s^{11}(1-\mathbb{F}_s^1(\theta_s(\alpha_a,\alpha_b))) + T_s^{10}\mathbb{F}_s^1(\theta_s(\alpha_a,\alpha_b)))}{T_s^{01}(1-\mathbb{F}_s^0(\theta_s(\alpha_a,\alpha_b))) + T_s^{00}\mathbb{F}_s^0(\theta_s(\alpha_a,\alpha_b))}.$$

$\forall \alpha_a, \alpha_b \in [0,1]$, increasing any T_s^{yd} decreases $\frac{1-g_s^1(\theta_s(\alpha_a,\alpha_b))}{g_s^0(\theta_s(\alpha_a,\alpha_b))}$. Let $\psi_{s'}(\cdot)$ be the consequent balanced function after increasing T_s^{yd}, and $\widehat{\alpha}_{s'}$ be corresponding equilibrium. Given any $\alpha_a, \alpha_b \in [0,1]$, we have $\psi_a(\alpha_b) < \psi_{a'}(\alpha_b)$ and $\psi_b(\alpha_a) < \psi_{b'}(\alpha_a)$. Therefore, $\widehat{\alpha}_{a'} > \widehat{\alpha}_a$ and $\widehat{\alpha}_{b'} > \widehat{\alpha}_b$.

APPENDIX E

Impact of Fairness Interventions on Strategic Manipulation

E.1 Generalization to high dimensional feature space

All analysis and conclusions can be generalized to high dimensional feature space $X \in \mathbb{R}^d$. In this case, high dimensional features are first mapped to one dimensional qualification profile $\gamma_s(x) = P_{Y|X,S}(1|x,s)$, based on which the decision maker makes decisions about individuals. A threshold policy is in the form of $\pi_s(x) = \mathbf{1}(\gamma_s(x) \geq \phi_s)$ with threshold $\phi_s \in [0,1]$.

Let $\gamma_s^{-1}(l_s) \subset \mathbb{R}^b$ be defined as the preimage of l_s under qualification profile γ_s, then we can adjust all analysis using $\gamma_s^{-1}(\cdot)$. For example, the strict monotone likelihood ratio property in Assumption 11 can be adjusted as follows: *for any two likelihoods* $0 \leq \underline{l_s} < \overline{l_s} \leq 1$*, we have* $\gamma_s^{-1}([\overline{l_s}, 1]) \subset \gamma_s^{-1}([\underline{l_s}, 1])$, i.e., any individual who can get accepted under threshold $\overline{l_s}$ can also be accepted under any lower threshold $\underline{l_s}$.

Because $\gamma_s(x) = P_{Y|X,S}(1|x,s) = \frac{1}{1+\frac{f_s^0(x)}{f_s^0(1)}\frac{(1-\alpha_s)}{\alpha_s}}$, (non-)strategic (fair) threshold ϕ_s in the space of qualification profile can be found based on $\frac{f_s^1(\theta_s)}{f_s^0(\theta_s)}$ given in Lemmas 8-11. Specifically, replace $\frac{f_s^1(\theta_s)}{f_s^0(\theta_s)}$ with $\frac{1-\alpha_s}{\alpha_s}\frac{\phi_s}{1-\phi_s}$, and $\Delta_s(\theta_s)$ with $\int_{x\in\gamma_s^{-1}([\phi_s,1])} f_s^0(1) - f_s^0(x)dx$ in Lemmas 8-11. Then the consequent policy $\pi_s(x) = \mathbf{1}(\gamma_s(x) \geq \phi_s)$ is the optimal policy.

E.2 Proof that non-strategic policy is threshold policy

The non-strategic optimal policy $\widehat{\pi}_s^{\text{UN}} = \arg\max_{\pi_s} \widehat{U}_s(\pi_s)$ is given by

$$\widehat{\pi}_s^{\text{UN}}(x) = \begin{cases} 1, \text{ if } \frac{f_s^0(1)}{f_s^0(x)} \geq \frac{u_-(1-\alpha_s)}{u_+\alpha_s} \\ 0, \text{ o.w.} \end{cases} \tag{E.1}$$

Re-writing based on qualification the profile $\gamma_s(x) = \dfrac{1}{\frac{f_s^0(x)}{f_s^0(1)}\frac{(1-\alpha_s)}{\alpha_s} + 1}$, (E.1) is reduced to

$$\widehat{\pi}_s^{\text{UN}}(x) = \mathbf{1}\Big(\gamma_s(x) \geq \frac{u_-}{u_+ + u_-}\Big).$$

E.3 Proof of Lemma 8

Let $\pi_s(x) = \mathbf{1}(x \geq \theta)$, then $\widehat{U}_s(\pi_s) := \widehat{U}_s(\theta)$ can be written as

$$\begin{aligned} \widehat{U}_s(\theta) &= u_+\alpha_s(1-\mathbb{F}_s^1(\theta)) - u_-(1-\alpha_s)(1-\mathbb{F}_s^0(\theta)) \\ &= u_+\alpha_s - u_-(1-\alpha_s) + u_-(1-\alpha_s)\mathbb{F}_s^0(\theta) - u_+\alpha_s\mathbb{F}_s^1(\theta) \\ \frac{\partial \widehat{U}_s(\theta)}{\partial \theta} &= u_-(1-\alpha_s)f_s^0(\theta) - u_+\alpha_s f_s^1(\theta) \end{aligned}$$

Under Assumption 11, $\widehat{U}_s(\theta)$ increases over $\theta \leq \widehat{\theta}_s^{\text{UN}}$ and decreases over $\theta \geq \widehat{\theta}_s^{\text{UN}}$. $\widehat{\theta}_s^{\text{UN}}$ is the optimal threshold and is the unique extreme point of $\widehat{U}_s(\theta)$.

E.4 Deviation of Manipulation Probability

When $\pi_s(x) = \mathbf{1}(x \geq \theta)$ is a threshold policy, we have

$$\begin{aligned} P_{D|Y,M,S}(1|y,m,s) &= \int_X P_{D,X|Y,M,S}(1,x|y,m,s)dx \\ &= \int_X P_{D|X,Y,M,S}(1|x,y,m,s)P_{X|Y,M,S}(x|y,m,s)dx \\ &= \int_X \pi_s(x)P_{X|Y,M,S}(x|y,m,s)dx = 1 - \mathbb{F}_{X|Y,M,S}(\theta|y,m,s) \end{aligned}$$

Therefore,

$$\begin{aligned} p_s^0(\pi_s) &= \mathbb{F}_{C_s}\Big(P_{D|Y,M,S}(1|0,1,s) - P_{D|Y,M,S}(1|0,0,s)\Big) \\ &= \mathbb{F}_{C_s}\Big(\mathbb{F}_{X|Y,M,S}(\theta|0,0,s) - \mathbb{F}_{X|Y,M,S}(\theta|0,1,s)\Big) = \mathbb{F}_{C_s}\Big(\mathbb{F}_s^0(\theta) - \mathbb{F}_s^1(\theta)\Big). \end{aligned}$$

E.5 Proof of Lemma 10

Take derivative of $U_s(\theta)$ w.r.t. θ, we have

$$\begin{aligned} \frac{\partial U_s(\theta)}{\partial \theta} &= \Big(f_s^0(\theta)(u_-(1-\alpha_s) - \Psi_s'(\Delta_s(\theta))) + f_s^1(\theta)\Psi_s'(\Delta_s(\theta))\Big) - u_+\alpha_s f_s^1(\theta) \\ &\propto \Big(\frac{f_s^0(\theta)}{f_s^1(\theta)}(u_-(1-\alpha_s) - \Psi_s'(\Delta_s(\theta))) + \Psi_s'(\Delta_s(\theta))\Big) - u_+\alpha_s \end{aligned}$$

As $\theta \to \pm\infty$, $\Delta_s(\theta) \to 0$, $\Psi_s'(\Delta_s(\theta)) \to 0$ and $\frac{\partial U_s(\theta)}{\partial \theta} \propto u_-(1-\alpha_s)\frac{f_s^0(\theta)}{f_s^1(\theta)} - u_+\alpha_s$. Therefore, $\frac{\partial U_s(\theta)}{\partial \theta} > 0$ as $\theta \to -\infty$ and $\frac{\partial U_s(\theta)}{\partial \theta} < 0$ as $\theta \to +\infty$.

The strategic optimal threshold θ_s^{UN} satisfies

$$\frac{f_s^0(\theta_s^{\text{UN}})}{f_s^1(\theta_s^{\text{UN}})} = \frac{u_+\alpha_s - \Psi_s'(\Delta_s(\theta_s^{\text{UN}}))}{u_-(1-\alpha_s) - \Psi_s'(\Delta_s(\theta_s^{\text{UN}}))}.$$

E.6 Proof of Lemma 11

To satisfy fairness constraint $\mathcal{C}$, $\int_{\theta_a}^{\infty} \mathcal{P}_a^{\mathcal{C}}(x)dx = \int_{\theta_b}^{\infty} \mathcal{P}_b^{\mathcal{C}}(x)dx$ should hold. Denote CDF $\mathbb{F}_s^{\mathcal{C}}(\theta_s) = \int_{-\infty}^{\theta_s} \mathcal{P}_s^{\mathcal{C}}(x)dx$, then for any pair (θ_a, θ_b) that is fair, we have $\theta_a = (\mathbb{F}_a^{\mathcal{C}})^{-1}\mathbb{F}_b^{\mathcal{C}}(\theta_b) = \eta^{\mathcal{C}}(\theta_b)$ hold for some strictly increasing function $\eta^{\mathcal{C}}(\cdot)$. Denote $u = \mathbb{F}_b^{\mathcal{C}}(\theta_b)$ and $\theta_a = (\mathbb{F}_a^{\mathcal{C}})^{-1}(u)$, the following holds:

$$\frac{d\eta^{\mathcal{C}}(\theta_b)}{d\theta_b} = \frac{d(\mathbb{F}_a^{\mathcal{C}})^{-1}\mathbb{F}_b^{\mathcal{C}}(\theta_b)}{d\theta_b} = \frac{d(\mathbb{F}_a^{\mathcal{C}})^{-1}(u)}{du}\frac{du}{d\theta_b} = \frac{1}{(\mathbb{F}_a^{\mathcal{C}})'((\mathbb{F}_a^{\mathcal{C}})^{-1}(\theta_b))}\frac{du}{d\theta_b} = \frac{(\mathbb{F}_b^{\mathcal{C}})'(\theta_b)}{(\mathbb{F}_a^{\mathcal{C}})'(\theta_a)} = \frac{\mathcal{P}_b^{\mathcal{C}}(\theta_b)}{\mathcal{P}_a^{\mathcal{C}}(\theta_a)}$$

The total utility can be written as a function of θ_b, take the derivative of $n_a U_a(\eta^{\mathcal{C}}(\theta_b)) + n_b U_b(\theta_b)$

w.r.t. θ_b, the optimal $\theta_b^{\mathcal{C}}$ satisfies the following,

$$n_a \frac{dU_a(\eta^{\mathcal{C}}(\theta_b))}{d\theta_b}\Big|_{\theta_b=\theta_b^{\mathcal{C}}} \frac{d\eta^{\mathcal{C}}(\theta_b)}{d\theta_b}\Big|_{\theta_b=\theta_b^{\mathcal{C}}} + n_b \frac{dU_b(\theta_b)}{d\theta_b}\Big|_{\theta_b=\theta_b^{\mathcal{C}}} = 0$$
$$\iff n_a \frac{dU_a(\eta^{\mathcal{C}}(\theta_b))}{d\theta_b}\Big|_{\theta_b=\theta_b^{\mathcal{C}}} \frac{\mathcal{P}_b^{\mathcal{C}}(\theta_b^{\mathcal{C}})}{\mathcal{P}_a^{\mathcal{C}}(\eta^{\mathcal{C}}(\theta_b^{\mathcal{C}}))} + n_b \frac{dU_b(\theta_b)}{d\theta_b}\Big|_{\theta_b=\theta_b^{\mathcal{C}}} = 0$$

Simplifying above equation gives the result.

E.7 Proof of Theorem 22

According to Lemma **??**, $(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ satisfies

$$\frac{f_s^0(\theta_s^{\text{UN}})}{f_s^1(\theta_s^{\text{UN}})} = \frac{u_+\alpha_s - \Psi_s'(\Delta_s(\theta_s^{\text{UN}}))}{u_-(1-\alpha_s) - \Psi_s'(\Delta_s(\theta_s^{\text{UN}}))} := \Omega_s(\theta_s^{\text{UN}})$$

Under Assumption 11, $\Delta_s(\theta)$ is single-peaked with maximum occurring at x_s^*. Define function $\Omega_s(\theta) := \frac{u_+\alpha_s - \Psi_s'(\Delta_s(\theta))}{u_-(1-\alpha_s) - \Psi_s'(\Delta_s(\theta))}$.

1. If $\alpha_s = \delta_u$, then

$$\frac{\partial U_s(\theta)}{\partial \theta} \propto \Big(\frac{f_s^0(\theta)}{f_s^1(\theta)} - 1\Big)(u_+\alpha_s - \Psi_s'(\Delta_s(\theta)))$$

   consider two cases:

   - $\overline{\Psi_s'} \le u_-(1-\alpha_s)$
     $\theta_s^{\text{UN}} = \widehat{\theta}_s^{\text{UN}} = x_s^*$ is unique optimal solution.
   - $\overline{\Psi_s'} > u_-(1-\alpha_s)$
     $U_s(\theta)$ has three extreme points where both $\theta_s^{\text{UN}} = \overline{x_s}$, $\theta_s^{\text{UN}} = \underline{x_s}$ are optimal, and x_s^* is the other extreme point that is not optimal.

2. If $\alpha_s < \delta_u$, then consider two cases:

   - $\overline{\Psi_s'} \le u_-(1-\alpha_s)$

$\Omega_s(\theta)$ decreases over $\theta < x_s^*$ and increases over $\theta > x_s^*$. $\Omega_s(\theta) \to \frac{u_+\alpha_s}{u_-(1-\alpha_s)} < 1$ as $\theta \to \pm\infty$. Under Assumption 11, $\frac{f_s^0(\theta)}{f_s^1(\theta)}$ intersects with $\Omega_s(\theta)$ at one unique point, i.e., θ_s^{UN} is unique and satisfies $\theta_s^{\text{UN}} > \widehat{\theta}_s^{\text{UN}} > x_s^*$.

- $\overline{\Psi'_s} > u_-(1-\alpha_s)$

  $\Omega_s(\theta)$ decreases from $\frac{u_+\alpha_s}{u_-(1-\alpha_s)}$ to $-\infty$ over $\theta < \underline{x_s}$; increases from $-\infty$ to $\frac{u_+\alpha_s}{u_-(1-\alpha_s)}$ over $\theta > \overline{x_s}$; decreases over $\theta \in (\underline{x_s}, x_s^*)$ and increases over $\theta \in (x_s^*, \overline{x_s})$.

  Because $\frac{f_s^0(x_s^*)}{f_s^1(x_s^*)} = 1$ and $\Omega_s(x_s^*) = 1 + \frac{u_-(1-\alpha_s)-u_+\alpha_s}{\overline{\Psi'_s}-u_-(1-\alpha_s)} > 1$, under Assumption 11, there exists a unique $\theta_s^{\text{UN}} > \widehat{\theta}_s^{\text{UN}} > x_s^*$ at which $\frac{f_s^0(\theta)}{f_s^1(\theta)}$ intersects with $\Omega_s(\theta)$, and $\theta_s^{\text{UN}} > \overline{x_s}$.

  Moreover, if $\exists\ \theta$ s.t. $\frac{f_s^0(\theta)}{f_s^1(\theta)} > \Omega_s(\theta)$, then $\frac{f_s^0(\theta)}{f_s^1(\theta)}$ will also intersects with $\Omega_s(\theta)$ at least two more points over $(\underline{x_s}, x_s^*)$.

  Next, we show that among all the extreme points, the one satisfying $\theta_s^{\text{UN}} > x_s^*$ is the optimal.

  Re-organize $U_s(\theta)$, we have

$$\arg\max_\theta U_s(\theta) = \arg\max_\theta \underbrace{\Delta_s(\theta)(1-\mathbb{F}_{C_s}(\Delta_s(\theta)))}_{:=h_1(\theta)} + \underbrace{\mathbb{F}_s^1(\theta)\Big(1-\frac{u_+\alpha_s}{u_-(1-\alpha_s)}\Big)}_{:=h_2(\theta)}$$

  For any extreme point $\theta' \in (\underline{x_s}, x_s^*)$, always there exists a point $x' > x_s^*$ satisfying $\Delta_s(x') = \Delta_s(\theta')$, so that $h_1(x') = h_1(\theta')$. Since $x' > \theta'$, $h_2(x') > h_2(\theta')$ holds so that $U_s(x') > U_s(\theta')$. In other words, $\exists$ a point over $(x_s^*, \overline{x_s})$ whose utility is higher than those of extreme points in $(\underline{x_s}, x_s^*)$. Since θ_s^{UN} is the optimal over $(x_s^*, \overline{x_s})$. It implies that θ_s^{UN} is optimal.

3. If $\alpha_s > \delta_u$, then consider two cases:

   - $\overline{\Psi'_s} \le u_+\alpha_s$

     $\frac{1}{\Omega_s(\theta)}$ decreases over $\theta < x_s^*$ and increases over $\theta > x_s^*$. $\frac{1}{\Omega_s(\theta)} \to \frac{u_-(1-\alpha_s)}{u_+\alpha_s} < 1$ as $\theta \to \pm\infty$. Under Assumption 11, $\frac{f_s^1(\theta)}{f_s^0(\theta)}$ intersects with $\frac{1}{\Omega_s(\theta)}$ at one unique point, i.e., θ_s^{UN} is unique and satisfies $\theta_s^{\text{UN}} < \widehat{\theta}_s^{\text{UN}} < x_s^*$.

   - $\overline{\Psi'_s} > u_+\alpha_s$

     $\frac{1}{\Omega_s(\theta)}$ decreases from $\frac{u_-(1-\alpha_s)}{u_+\alpha_s}$ to $-\infty$ over $\theta < \underline{z_s}$; increases from $-\infty$ to $\frac{u_-(1-\alpha_s)}{u_+\alpha_s}$ over $\theta > \overline{z_s}$; decreases over $\theta \in (\underline{z_s}, x_s^*)$ and increases over $\theta \in (x_s^*, \overline{z_s})$.

Because $\frac{f_s^1(x_s^*)}{f_s^0(x_s^*)} = 1$ and $\frac{1}{\Omega_s(\theta)} = 1 + \frac{u_+\alpha_s - u_-(1-\alpha_s)}{\overline{\Psi}'_s - u_+\alpha_s} > 1$, under Assumption 11, there exists a unique $\theta_s^{\text{UN}} < \widehat{\theta}_s^{\text{UN}} < x_s^*$ at which $\frac{f_s^0(\theta)}{f_s^1(\theta)}$ intersects with $\Omega_s(\theta)$, and $\theta_s^{\text{UN}} < \underline{z_s}$.

Moreover, if $\exists\ \theta$ s.t. $\frac{f_s^0(\theta)}{f_s^1(\theta)} < \Omega_s(\theta)$, then $\frac{f_s^0(\theta)}{f_s^1(\theta)}$ will also intersect with $\Omega_s(\theta)$ at least two more points over $(x_s^*, \overline{z_s})$.

We show that among all the extreme points, the one satisfying $\theta_s^{\text{UN}} < x_s^*$ is the optimal.

For any extreme point $\theta' \in (x_s^*, \overline{z_s})$, always there exists a point $x' < x_s^*$ satisfying $\Delta_s(x') = \Delta_s(\theta')$, so that $h_1(x') = h_1(\theta')$. Since $x' < \theta'$ and $1 < \frac{u_+\alpha_s}{u_-(1-\alpha_s)}$, $h_2(x') > h_2(\theta')$ holds so that $U_s(x') > U_s(\theta')$. In other words, $\exists$ a point over $(\underline{z_s}, x_s^*)$ whose utility is higher than those of extreme points in $(x_s^*, \overline{z_s})$. Since θ_s^{UN} is optimal over $(\underline{z_s}, x_s^*)$, it implies that θ_s^{UN} is optimal.

E.8 Proof of Theorem 23

WLOG, let $s := a$ and $-s := b$.

Because $\alpha_a > \delta_u > \alpha_b$, according to Theorem 22, we have $x_b^* < \widehat{\theta}_b^{\text{UN}} < \theta_b^{\text{UN}}$ and $x_a^* > \widehat{\theta}_a^{\text{UN}} > \theta_a^{\text{UN}}$. It implies that $\mathbb{F}_a^{\mathcal{C}}(x_a^*) > \mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}) > \mathbb{F}_a^{\mathcal{C}}(\theta_a^{\text{UN}})$ and $\mathbb{F}_b^{\mathcal{C}}(x_b^*) < \mathbb{F}_b^{\mathcal{C}}(\widehat{\theta}_b^{\text{UN}}) < \mathbb{F}_b^{\mathcal{C}}(\theta_b^{\text{UN}})$.

Since $\mathbb{F}_a^{\mathcal{C}}(x_a^*) \le \mathbb{F}_b^{\mathcal{C}}(x_b^*)$, we have $\mathbb{F}_a^{\mathcal{C}}(\theta_a^{\text{UN}}) < \mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}) < \mathbb{F}_b^{\mathcal{C}}(\widehat{\theta}_b^{\text{UN}}) < \mathbb{F}_b^{\mathcal{C}}(\theta_b^{\text{UN}})$, so that $\mathcal{E}^{\mathcal{C}}(\theta_a^{\text{UN}}, \theta_b^{\text{UN}}) > \mathcal{E}^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}}) > 0$.

E.9 Proof of Theorem 24

WLOG, let $s := a$ and $-s := b$.

By Theorem 22, $\theta_a^{\text{UN}} > \widehat{\theta}_a^{\text{UN}}$ always hold. If marginal manipulation gain of $\mathcal{G}_a$ is sufficiently small such that $\Psi'_a(\Delta_a(\widehat{\theta}_a^{\text{UN}})) \to 0$, then $\theta_a^{\text{UN}} \to \widehat{\theta}_a^{\text{UN}}$; If marginal manipulation gain of $\mathcal{G}_a$ is sufficiently large such that $\Psi'_a(\Delta_a(\widehat{\theta}_a^{\text{UN}})) \to u_-(1-\alpha_a)$, then $\theta_a^{\text{UN}} \gg \widehat{\theta}_a^{\text{UN}}$.

For any given $\mathcal{G}_b$, $\mathbb{F}_b^{\mathcal{C}}(\theta_b^{\text{UN}}) > \mathbb{F}_b^{\mathcal{C}}(\widehat{\theta}_b^{\text{UN}}) > \mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}})$, since any $\mathbb{F}_a^{\mathcal{C}}(\theta_a^{\text{UN}}) \in (\mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}), 1)$ is attainable by controlling manipulation cost C_a, it implies that there exists C_a s.t. $\left|\mathcal{E}^{\mathcal{C}}(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})\right| < \left|\mathcal{E}^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})\right|$ or $\mathbb{F}_b^{\mathcal{C}}(\theta_b^{\text{UN}}) < \mathbb{F}_a^{\mathcal{C}}(\theta_a^{\text{UN}})$.

E.10 Proof of Theorem 25

WLOG, let $s := a$ and $-s := b$.

1. $\alpha_a < \delta_u < \alpha_b$ and $\mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}) < \mathbb{F}_b^{\mathcal{C}}(\widehat{\theta}_b^{\text{UN}})$.

   Since $\alpha_a < \delta_u < \alpha_b$, we have $\widehat{\theta}_a^{\text{UN}} > x_a^*$ and $\widehat{\theta}_b^{\text{UN}} < x_b^*$.

   Under Assumption 11, $\widehat{U}_s(\theta)$ is non-decreasing over $(-\infty, \widehat{\theta}_s^{\text{UN}})$ and non-increasing over $(\widehat{\theta}_s^{\text{UN}}, +\infty)$. One of the followings must hold: (1) $\widehat{\theta}_a^{\mathcal{C}} > \widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\mathcal{C}} < \widehat{\theta}_b^{\text{UN}}$ (2) $\widehat{\theta}_a^{\mathcal{C}} < \widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\mathcal{C}} > \widehat{\theta}_b^{\text{UN}}$. Because if $\widehat{\theta}_a^{\mathcal{C}} > \widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\mathcal{C}} > \widehat{\theta}_b^{\text{UN}}$ or $\widehat{\theta}_a^{\mathcal{C}} < \widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\mathcal{C}} < \widehat{\theta}_b^{\text{UN}}$ holds, we can always find another pair of thresholds satisfying fairness $\mathcal{C}$ but achieves a higher utility $\sum_{s=a,b} n_s \widehat{U}_s(\theta_s)$ so that $(\widehat{\theta}_a^{\mathcal{C}}, \widehat{\theta}_b^{\mathcal{C}})$ cannot be non-strategic optimal fair policy.

   Because $\mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}) < \mathbb{F}_b^{\mathcal{C}}(\widehat{\theta}_b^{\text{UN}})$ and $\mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\mathcal{C}}) = \mathbb{F}_b^{\mathcal{C}}(\widehat{\theta}_b^{\mathcal{C}})$, $\widehat{\theta}_a^{\mathcal{C}} > \widehat{\theta}_a^{\text{UN}} > x_a^*, \widehat{\theta}_b^{\mathcal{C}} < \widehat{\theta}_b^{\text{UN}} < x_b^*$ must hold.

   If $\Psi_a'(\Delta_a(\widehat{\theta}_a^{\mathcal{C}})) > u_-(1-\alpha_a)$ and $\Psi_b'(\Delta_b(\widehat{\theta}_b^{\mathcal{C}})) > u_+\alpha_b$, then we have $\widehat{\theta}_a^{\mathcal{C}} < \overline{x_a}$ and $\widehat{\theta}_b^{\mathcal{C}} > \underline{z_b}$, where $\overline{x_a}, \underline{z_b}$ are defined s.t. $\Psi_a'(\Delta_a(\overline{x_a})) = u_-(1-\alpha_a)$ and $\Psi_b'(\Delta_b(\underline{z_b})) = u_+\alpha_b$. By Theorem 22, $U_a(\theta)$ is increasing over $(x_a^*, \overline{x_a})$ and $U_b(\theta)$ is decreasing over $(\underline{z_b}, x_b^*)$. It implies that $U_a(\widehat{\theta}_a^{\mathcal{C}}) > U_a(\widehat{\theta}_a^{\text{UN}})$ and $U_b(\widehat{\theta}_b^{\mathcal{C}}) > U_b(\widehat{\theta}_b^{\text{UN}})$.

2. $\alpha_a, \alpha_b > \delta_u$, $\mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}) < \mathbb{F}_b^{\mathcal{C}}(\widehat{\theta}_b^{\text{UN}})$, and $\alpha_a \to \delta_u$.

   Since $\alpha_a, \alpha_b > \delta_u$, we have $\widehat{\theta}_a^{\text{UN}} < x_a^*$ and $\widehat{\theta}_b^{\text{UN}} < x_b^*$.

   Because $\mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}) < \mathbb{F}_b^{\mathcal{C}}(\widehat{\theta}_b^{\text{UN}})$ and $\mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\mathcal{C}}) = \mathbb{F}_b^{\mathcal{C}}(\widehat{\theta}_b^{\mathcal{C}})$, $\widehat{\theta}_a^{\mathcal{C}} > \widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\mathcal{C}} < \widehat{\theta}_b^{\text{UN}}$ must hold.

   If $\alpha_a \to \delta_u$, then $\widehat{\theta}_a^{\text{UN}} \to x_a^*$ and $\widehat{\theta}_a^{\text{UN}} < x_a^* < \widehat{\theta}_a^{\mathcal{C}}$ hold.

   If $\Psi_a'(\Delta_a(\widehat{\theta}_a^{\mathcal{C}})) > u_+\alpha_a$, $\Psi_b'(\Delta_b(\widehat{\theta}_b^{\mathcal{C}})) > u_+\alpha_b$, then we have $\widehat{\theta}_a^{\mathcal{C}} < \overline{z_a}$ and $\widehat{\theta}_b^{\mathcal{C}} > \underline{z_b}$. By Theorem 22, $U_b(\theta)$ is decreasing over $(\underline{z_b}, x_b^*)$ implying $U_b(\widehat{\theta}_b^{\mathcal{C}}) > U_b(\widehat{\theta}_b^{\text{UN}})$, and $U_a(\theta)$ may have additional extreme points over $(x_a^*, \overline{z_a})$. Specifically, as $\alpha_a \to \delta_u$, there are two extreme points x_1, x_2 with $x_1 \to x_a^*$, $x_2 \to \overline{z_a}$ (by Theorem 22), Because $U_a(\theta)$ is increasing over $[x_1, x_2]$, $U_a(x_2) \to U_a(\theta_a^{\text{UN}}) = \max_\theta U_a(\theta)$, and $U_a(x_1) \to U_a(x_a^*), U_a(\widehat{\theta}_a^{\text{UN}}) \to U_a(x_a^*)$, we have $U_a(\widehat{\theta}_a^{\mathcal{C}}) > U_a(\widehat{\theta}_a^{\text{UN}})$.

3. $\alpha_a, \alpha_b < \delta_u$, $\mathbb{F}_a^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}) < \mathbb{F}_b^{\mathcal{C}}(\widehat{\theta}_b^{\text{UN}})$, and $\alpha_b \to \delta_u$.

   It can be proved similarly as case 2 and is omitted.

E.11 Proof of Theorem 26

For any pair (θ_a, θ_b) satisfying fairness $\mathcal{C}$, $\mathbb{F}_a^{\mathcal{C}}(\theta_a) = \mathbb{F}_b^{\mathcal{C}}(\theta_b)$ should hold. We have $\theta_a = (\mathbb{F}_a^{\mathcal{C}})^{-1}\mathbb{F}_b^{\mathcal{C}}(\theta_b) = \eta^{\mathcal{C}}(\theta_b)$ for some strictly increasing function $\eta^{\mathcal{C}}(\cdot)$.

1. Both $U_a(\theta)$ and $U_b(\theta)$ have unique extreme points.

   Prove $\theta_a^{\text{UN}} > \theta_a^{\mathcal{C}}, \theta_b^{\text{UN}} < \theta_b^{\mathcal{C}}$ or $\theta_a^{\text{UN}} < \theta_a^{\mathcal{C}}, \theta_b^{\text{UN}} > \theta_b^{\mathcal{C}}$ by contradiction. Suppose $\theta_a^{\text{UN}} > \theta_a^{\mathcal{C}}, \theta_b^{\text{UN}} > \theta_b^{\mathcal{C}}$, then we can always find another pair of thresholds (θ_a', θ_b') that satisfies $\mathcal{C}$ with $\theta_a^{\mathcal{C}} < \theta_a' \leq \theta_a^{\text{UN}}$ and $\theta_b^{\mathcal{C}} < \theta_b' \leq \theta_b^{\text{UN}}$. Because $U_s(\theta)$ has unique extreme point and it increases over $\theta < \theta_s^{\text{UN}}$, $U_s(\theta_s^{\mathcal{C}}) < U_s(\theta_s'), \forall s \in \{a, b\}$ holds, i.e., $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$ can not be the optimal pair that satisfies the fairness. Similarly, we can show that $\theta_a^{\text{UN}} < \theta_a^{\mathcal{C}}, \theta_b^{\text{UN}} < \theta_b^{\mathcal{C}}$ cannot hold.

   Let x_s^{UN} be defined s.t. $\Delta_s(x_s^{\text{UN}}) = \Delta_s(\theta_s^{\text{UN}})$ and $x_s^{\text{UN}} \neq \theta_s^{\text{UN}}$ when $\theta_s^{\text{UN}} \neq x_s^*$. Note that x_s^{UN} is the point at which $p_s^0(x_s^{\text{UN}}) = p_s^0(\theta_s^{\text{UN}})$. WLOG, let $s := a$ and $-s := b$.

   Let $x_a^{\mathcal{C}} := \eta^{\mathcal{C}}(x_b^{\text{UN}})$, i.e., $(x_a^{\mathcal{C}}, x_b^{\text{UN}})$ satisfies fairness constraint $\mathcal{C}$. Given any fixed α_b, as α_a changes, x_a^{UN}, $x_a^{\mathcal{C}}$, and θ_a^{UN} also change. Rewrite them as functions of α_a, i.e., $x_a^{\text{UN}}(\alpha_a)$, $x_a^{\mathcal{C}}(\alpha_a) := \eta^{\mathcal{C}}(x_b^{\text{UN}}; \alpha_a)$, and $\theta_a^{\text{UN}}(\alpha_a)$.

   - $\alpha_a > \delta_u > \alpha_b$

     $x_a^{\text{UN}}(\alpha_a)$ increases in $\alpha_a \in (\delta_u, 1)$

     $$\lim_{\alpha_a \to \delta_u} x_a^{\text{UN}}(\alpha_a) = x_a^*, \quad \lim_{\alpha_a \to 1} x_a^{\text{UN}}(\alpha_a) = +\infty$$

     $\theta_a^{\text{UN}}(\alpha_a)$ decreases in $\alpha_a \in (\delta_u, 1)$

     $$\lim_{\alpha_a \to \delta_u} \theta_a^{\text{UN}}(\alpha_a) = x_a^*, \quad \lim_{\alpha_a \to 1} \theta_a^{\text{UN}}(\alpha_a) = -\infty$$

     $x_a^{\mathcal{C}}(\alpha_a)$ is non-decreasing in α_a

     $$\lim_{\alpha_a \to \delta_u} x_a^{\mathcal{C}}(\alpha_a) = \eta^{\mathcal{C}}(x_b^{\text{UN}}; \delta_u) < +\infty, \quad \lim_{\alpha_a \to 1} x_a^{\mathcal{C}}(\alpha_a) = \eta^{\mathcal{C}}(x_b^{\text{UN}}; 1) < +\infty$$

     Therefore, $\exists\, \kappa > \delta_u$ s.t. for any $\alpha_a > \kappa$, $x_a^{\mathcal{C}}(\alpha_a) \in (\theta_a^{\text{UN}}(\alpha_a), x_a^{\text{UN}}(\alpha_a))$.

As $n_a \to 1$, $\theta_a^{\mathcal{C}} \to \theta_a^{\text{UN}}$. Therefore, $\forall \alpha_a \in (\kappa, 1)$, there exists $\tau \in (0,1)$ s.t. $\forall n_a > \tau$, we have $\theta_a^{\mathcal{C}} \in (\theta_a^{\text{UN}}, x_a^{\mathcal{C}})$ and $\theta_b^{\mathcal{C}} < x_b^{\text{UN}}$. It implies that $\Delta_a(\theta_a^{\mathcal{C}}) > \Delta_a(\theta_a^{\text{UN}})$ and $\Delta_b(\theta_b^{\mathcal{C}}) < \Delta_a(\theta_b^{\text{UN}})$ so that $p_a^{\mathcal{C}} > p_a^{\text{UN}}$ and $p_b^{\mathcal{C}} < p_b^{\text{UN}}$.

- $\alpha_a, \alpha_b > \delta_u$

  From the above, $\exists \kappa > \delta_u$ s.t. $\forall \alpha_a > \kappa$, $x_a^{\mathcal{C}}(\alpha_a) \in (\theta_a^{\text{UN}}(\alpha_a), x_a^{\text{UN}}(\alpha_a))$.

  Since $U_a(\theta)$, $U_b(\theta)$ have unique extreme points, neither $\theta_a^{\mathcal{C}} > \theta_a^{\text{UN}}, \theta_b^{\mathcal{C}} > \theta_b^{\text{UN}}$ nor $\theta_a^{\mathcal{C}} < \theta_a^{\text{UN}}, \theta_b^{\mathcal{C}} < \theta_b^{\text{UN}}$ hold. When $\alpha_a > \kappa$, either of the followings holds: (1) $\theta_a^{\mathcal{C}} < \theta_a^{\text{UN}}$, $\theta_b^{\mathcal{C}} \in (\theta_b^{\text{UN}}, x_b^{\text{UN}})$; (2) $\theta_b^{\mathcal{C}} < \theta_b^{\text{UN}}$, $\theta_a^{\mathcal{C}} \in (\theta_a^{\text{UN}}, x_a^{\mathcal{C}})$. It implies $p_b^{\mathcal{C}} > p_b^{\text{UN}}$, $p_a^{\mathcal{C}} < p_a^{\text{UN}}$, or $p_a^{\mathcal{C}} > p_a^{\text{UN}}$, $p_b^{\mathcal{C}} < p_b^{\text{UN}}$.

- $\alpha_a, \alpha_b < \delta_u$

  Prove in the similar way. $\exists \kappa < \delta_u$ s.t. $\forall \alpha_a < \kappa$, $x_a^{\mathcal{C}}(\alpha_a) \in (x_a^{\text{UN}}(\alpha_a), \theta_a^{\text{UN}}(\alpha_a))$.

  Since $U_a(\theta)$, $U_b(\theta)$ have unique extreme points, either of the followings holds when $\alpha_a < \kappa$: (1) $\theta_a^{\mathcal{C}} > \theta_a^{\text{UN}}$, $\theta_b^{\mathcal{C}} \in (x_b^{\text{UN}}, \theta_b^{\text{UN}})$; (2) $\theta_b^{\mathcal{C}} > \theta_b^{\text{UN}}$, $\theta_a^{\mathcal{C}} \in (x_a^{\mathcal{C}}, \theta_a^{\text{UN}})$. It implies $p_a^{\mathcal{C}} < p_a^{\text{UN}}$, $p_b^{\mathcal{C}} > p_b^{\text{UN}}$, or $p_b^{\mathcal{C}} < p_b^{\text{UN}}$, $p_a^{\mathcal{C}} > p_a^{\text{UN}}$.

2. At least one of $U_a(\theta)$, $U_b(\theta)$ has multiple extreme points. WLOG, let $s := a$ and $-s := b$.

   - $\alpha_a > \delta_u > \alpha_b$

     (i) $U_a(\theta)$ has multiple extreme points while $U_b(\theta)$ has a unique extreme point.
     Let x_1, x_2 be two extreme points over $(x_a^*, \overline{z_a})$ with x_2 being the optimal extreme point over $(x_a^*, \overline{z_a})$ and x_1 the largest extreme point satisfying $x_1 < x_2$. By Theorem 22, $\theta_a^{\text{UN}} < x_a^*$.
     As $n_b \to 1$, $\theta_b^{\mathcal{C}} \to \theta_b^{\text{UN}}$ and $\theta_a^{\mathcal{C}} \to \eta^{\mathcal{C}}(\theta_b^{\text{UN}})$. If $\eta^{\mathcal{C}}(\theta_b^{\text{UN}}) \in (x_1, x_2)$ happens to be satisfied under groups' feature distributions and manipulation costs, then it's possible that there exists a sufficiently large n_b such that the a fair threshold pair $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$ results in a higher total utility than that of $(\eta^{\mathcal{C}}(\theta_b^{\text{UN}}), \theta_b^{\text{UN}})$. In this case, $\theta_a^{\mathcal{C}} > \theta_a^{\text{UN}}, \theta_b^{\mathcal{C}} > \theta_b^{\text{UN}}$ and $\theta_a^{\mathcal{C}} \in (\eta^{\mathcal{C}}(\theta_b^{\text{UN}}), x_2)$ must hold.
     Because $\theta_a^{\text{UN}} < \underline{z_s}$, $\theta_a^{\mathcal{C}} < \overline{z_a}$, we have $\Delta_a(\theta_a^{\text{UN}}) < \Delta_a(\theta_a^{\mathcal{C}})$ and $p_a^{\mathcal{C}} > p_a^{\text{UN}}$.
     Because $\alpha_b < \delta_u$, we have $\theta_b^{\text{UN}} > x_b^*$. Since $\theta_b^{\mathcal{C}} > \theta_b^{\text{UN}}$, it holds that $p_b^{\mathcal{C}} < p_b^{\text{UN}}$.

     (ii) $U_a(\theta)$ has a unique extreme point while $U_b(\theta)$ has multiple extreme points.
     Similar to the reasoning in (i), let x_1, x_2 be two extreme points over $(\underline{x_b}, x_b^*)$ with x_1 being the optimal extreme point over $(\underline{x_b}, x_b^*)$ and x_2 the smallest extreme point satisfying $x_1 < x_2$.

If $(\eta^{\mathcal{C}})^{-1}(\theta_a^{\text{UN}}) \in (x_1, x_2)$ happens to be satisfied, then it's possible to find a sufficiently large n_a such that the fair pair $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$ results in a higher utility than that of $(\theta_a^{\text{UN}}, (\eta^{\mathcal{C}})^{-1}(\theta_a^{\text{UN}}))$. In this case, $\theta_a^{\mathcal{C}} < \theta_a^{\text{UN}}, \theta_b^{\mathcal{C}} < \theta_b^{\text{UN}}$ and $\theta_b^{\mathcal{C}} \in (x_1, (\eta^{\mathcal{C}})^{-1}(\theta_a^{\text{UN}}))$ must hold.
Because $\theta_a^{\mathcal{C}} < \theta_a^{\text{UN}} < x_a^*$ and $\theta_b^{\text{UN}} > \overline{x_b}$, $\theta_b^{\mathcal{C}} > \underline{x_b}$, we have $\Delta_a(\theta_a^{\text{UN}}) > \Delta_a(\theta_a^{\mathcal{C}})$ and $\Delta_b(\theta_b^{\text{UN}}) < \Delta_b(\theta_b^{\mathcal{C}})$. As such, $p_a^{\mathcal{C}} < p_a^{\text{UN}}$, $p_b^{\mathcal{C}} > p_b^{\text{UN}}$.

(iii) Both $U_a(\theta)$, $U_b(\theta)$ have multiple extreme points.
In this case, $\theta_a^{\text{UN}} < x_a^*$ and all other extreme points of $U_a(\theta)$ fall in $(x_a^*, \overline{z_a})$ with $\underline{z_a} > \theta_a^{\text{UN}}$; $\theta_b^{\text{UN}} > x_b^*$ and all other extreme points of $U_b(\theta)$ fall in $(\underline{x_b}, x_b^*)$ with $\overline{x_b} < \theta_b^{\text{UN}}$.
If $\theta_a^{\mathcal{C}} < \theta_a^{\text{UN}}, \theta_b^{\mathcal{C}} < \theta_b^{\text{UN}}$ happens to be satisfied, then $\theta_b^{\mathcal{C}} \in (\underline{x_b}, x_b^*)$ must hold. It implies that $\Delta_a(\theta_a^{\text{UN}}) > \Delta_a(\theta_a^{\mathcal{C}})$ and $\Delta_b(\theta_b^{\text{UN}}) < \Delta_b(\theta_b^{\mathcal{C}})$. As such, $p_a^{\mathcal{C}} < p_a^{\text{UN}}$, $p_b^{\mathcal{C}} > p_b^{\text{UN}}$.
Similarly, if $\theta_a^{\mathcal{C}} > \theta_a^{\text{UN}}, \theta_b^{\mathcal{C}} > \theta_b^{\text{UN}}$ happens to be satisfied, then $\theta_a^{\mathcal{C}} \in (x_a^*, \overline{z_a})$ must hold. It implies that $p_a^{\mathcal{C}} > p_a^{\text{UN}}$, $p_b^{\mathcal{C}} < p_b^{\text{UN}}$.

- $\alpha_a, \alpha_b > \delta_u$

  In this case, $\theta_a^{\text{UN}} < x_a^*, \theta_b^{\text{UN}} < x_b^*$ and $U_a(\theta)$ (or $U_b(\theta)$) increases over $\theta < \theta_a^{\text{UN}}$ (or $\theta < \theta_b^{\text{UN}}$). WLOG, let $\mathcal{G}_a$ has multiple extreme points, while $\mathcal{G}_b$ may or may not have multiple extreme points.

  Note that $\theta_a^{\mathcal{C}} < \theta_a^{\text{UN}}, \theta_b^{\mathcal{C}} < \theta_b^{\text{UN}}$ cannot hold, otherwise always there exists a fair threshold pair (θ_a', θ_b') with $\theta_a' \in (\theta_a^{\mathcal{C}}, \theta_a^{\text{UN}})$ and $\theta_b' \in (\theta_b^{\mathcal{C}}, \theta_b^{\text{UN}})$ whose utility is higher than that of $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$.

  In contrast, $\theta_a^{\mathcal{C}} > \theta_a^{\text{UN}}, \theta_b^{\mathcal{C}} > \theta_b^{\text{UN}}$ may hold. In this case, $\theta_a^{\mathcal{C}} \in (x_a^*, \overline{z_a})$ must hold, while either $\theta_b^{\mathcal{C}} < x_b^{\text{UN}}$ or $\theta_b^{\mathcal{C}} > x_b^{\text{UN}}$ holds.

  Therefore, $\Delta_a(\theta_a^{\text{UN}}) < \Delta_a(\theta_a^{\mathcal{C}})$ and $\Delta_b(\theta_b^{\text{UN}}) < \Delta_b(\theta_b^{\mathcal{C}})$ (or $\Delta_b(\theta_b^{\text{UN}}) > \Delta_b(\theta_b^{\mathcal{C}})$) must hold so that $p_a^{\mathcal{C}} > p_a^{\text{UN}}$, $p_b^{\mathcal{C}} > p_b^{\text{UN}}$ (or $p_b^{\mathcal{C}} < p_b^{\text{UN}}$).

  We can prove in a similar way for the case when $\alpha_a, \alpha_b < \delta_u$.

E.12 Proof of Theorem 27

First consider case when $\alpha_a, \alpha_b > \delta_u$.

WLOG, let $s := a$ and $-s := b$.

Define function $\eta^{\mathcal{C}}(\cdot) := (\mathbb{F}_a^{\mathcal{C}})^{-1}\mathbb{F}_b^{\mathcal{C}}(\cdot)$. If $\mathbb{F}_b^{\mathcal{C}}(x_b^{\text{UN}}) < \mathbb{F}_a^{\mathcal{C}}(x_a^*)$, then $\eta^{\mathcal{C}}(x_b^{\text{UN}}) < x_a^*$.

As $\alpha_a \to \delta_u$, $\theta_a^{\text{UN}} \to x_a^*$. As α_a decreases, $\eta^{\mathcal{C}}(x_b^{\text{UN}})$ is non-increasing (constant w.r.t. α_a for `EqOpt` and decreases for `DP`). $\exists\, \kappa > \delta_u$ s.t. when $\alpha_a = \kappa$, $\theta_a^{\text{UN}} = \eta^{\mathcal{C}}(x_b^{\text{UN}})$. Then $\forall \alpha_a < \kappa$, $\eta^{\mathcal{C}}(x_b^{\text{UN}}) < \theta_a^{\text{UN}}$.

As $n_a \to 1$, $\theta_a^{\mathcal{C}} \to \theta_a^{\mathrm{UN}}$ and $\lim_{n_a \to 1} \theta_b^{\mathcal{C}} > x_b^{\mathrm{UN}}$. Therefore, $\exists \tau \in (0,1)$ s.t. for any $n_a > \tau$, we have $\theta_a^{\mathcal{C}} \in (\eta^{\mathcal{C}}(x_b^{\mathrm{UN}}), \theta_a^{\mathrm{UN}})$ and $\theta_b^{\mathcal{C}} > x_b^{\mathrm{UN}}$. It implies that $p_a^{\mathcal{C}} < p_a^{\mathrm{UN}}, p_b^{\mathcal{C}} < p_b^{\mathrm{UN}}$.

For the case when $\alpha_a, \alpha_b < \delta_u$, it can be proved in a similar way and is omitted.

E.13 Proof of Proposition 6

WLOG, let $s := a$, $-s := b$.

Since $f_a^y(x) = f_b^y(x)$, denote $\Delta(\cdot) = \Delta_a(\cdot) = \Delta_b(\cdot)$.

By Lemma 10, for $s \in \{a, b\}$, $\widehat{\theta}_s^{\mathrm{UN}}$ satisfies $\frac{f_s^1(\widehat{\theta}_s^{\mathrm{UN}})}{f_s^0(\widehat{\theta}_s^{\mathrm{UN}})} = \frac{u_-(1-\alpha_s)}{u_+\alpha_s}$. Since $f_a^y(x) = f_b^y(x)$, $\alpha_b < \alpha_a < \delta_u$, $\frac{u_-(1-\alpha_b)}{u_+\alpha_b} > \frac{u_-(1-\alpha_a)}{u_+\alpha_a}$. Under Assumption 11, we have $\widehat{\theta}_a^{\mathrm{UN}} < \widehat{\theta}_b^{\mathrm{UN}}$.

It implies that $\mathbb{F}_a^1(\widehat{\theta}_a^{\mathrm{UN}}) < \mathbb{F}_b^1(\widehat{\theta}_b^{\mathrm{UN}})$, so that $\mathbb{F}_a^{\mathtt{EqOpt}}(\widehat{\theta}_a^{\mathrm{UN}}) < \mathbb{F}_b^{\mathtt{EqOpt}}(\widehat{\theta}_b^{\mathrm{UN}})$.

Note that $\mathbb{F}_s^{\mathtt{DP}}(\widehat{\theta}_s^{\mathrm{UN}}) = \alpha_s \mathbb{F}_s^1(\widehat{\theta}_s^{\mathrm{UN}}) + (1-\alpha_s)\mathbb{F}_s^0(\widehat{\theta}_s^{\mathrm{UN}})$. Since $\mathbb{F}_a^0(\widehat{\theta}_a^{\mathrm{UN}}) < \mathbb{F}_b^0(\widehat{\theta}_b^{\mathrm{UN}})$ and $\alpha_b < \alpha_a$, we have $\mathbb{F}_a^{\mathtt{DP}}(\widehat{\theta}_a^{\mathrm{UN}}) < \mathbb{F}_b^{\mathtt{DP}}(\widehat{\theta}_b^{\mathrm{UN}})$.

First, we show that the unfairness can be mitigated under some cost random variable C_a.

Given α_b, C_b, θ_b^{UN} is determined and satisfies $\frac{f_b^0(\theta_b^{\mathrm{UN}})}{f_b^1(\theta_b^{\mathrm{UN}})} = \frac{u_+\alpha_b - \Psi_b'(\Delta(\theta_b^{\mathrm{UN}}))}{u_-(1-\alpha_b) - \Psi_b'(\Delta(\theta_b^{\mathrm{UN}}))}$ (by Lemma 10), where $\Delta(\theta) = \mathbb{F}_b^0(\theta) - \mathbb{F}_b^1(\theta) = \mathbb{F}_a^0(\theta) - \mathbb{F}_a^1(\theta)$.

Given any $\alpha_a \in (\alpha_b, \delta_u)$, if $\mathcal{G}_a$'s cost C_a satisfies $\frac{u_+\alpha_a - \Psi_a'(\Delta(\theta_b^{\mathrm{UN}}))}{u_-(1-\alpha_a) - \Psi_a'(\Delta(\theta_b^{\mathrm{UN}}))} = \frac{u_+\alpha_b - \Psi_b'(\Delta(\theta_b^{\mathrm{UN}}))}{u_-(1-\alpha_b) - \Psi_b'(\Delta(\theta_b^{\mathrm{UN}}))}$, i.e.,

$$\Psi_a'(\Delta(\theta_b^{\mathrm{UN}})) = \underbrace{\frac{u_-(1-\alpha_a) - u_+\alpha_a}{u_-(1-\alpha_b) - u_+\alpha_b}}_{>0 \text{ (since } \alpha_a, \alpha_b < \delta_u)} \cdot \underbrace{(\Psi_b'(\Delta(\theta_b^{\mathrm{UN}})) - u_+\alpha_b)}_{<0 \text{ (by Theorem 22)}} + u_+\alpha_a < u_+\alpha_a < u_-(1-\alpha_a) \tag{E.2}$$

then $\frac{f_a^0(\theta_b^{\mathrm{UN}})}{f_a^1(\theta_b^{\mathrm{UN}})} = \frac{u_+\alpha_a - \Psi_a'(\Delta(\theta_b^{\mathrm{UN}}))}{u_-(1-\alpha_a) - \Psi_a'(\Delta(\theta_b^{\mathrm{UN}}))}$ holds and $\theta_a^{\mathrm{UN}} = \theta_b^{\mathrm{UN}}$.

Therefore, $\mathbb{F}_a^{\mathtt{EqOpt}}(\theta_a^{\mathrm{UN}}) = \mathbb{F}_{X|Y,S}(\theta_a^{\mathrm{UN}}|1,a) = \mathbb{F}_{X|Y,S}(\theta_b^{\mathrm{UN}}|1,b) = \mathbb{F}_b^{\mathtt{EqOpt}}(\theta_b^{\mathrm{UN}})$.

Because $\mathbb{F}_a^0(\theta_a^{\mathrm{UN}}) = \mathbb{F}_b^0(\theta_b^{\mathrm{UN}})$ also holds,

$$\begin{aligned}
\left|\mathbb{F}_a^{\mathtt{DP}}(\theta_a^{\mathrm{UN}}) - \mathbb{F}_b^{\mathtt{DP}}(\theta_b^{\mathrm{UN}})\right| &= (\alpha_a - \alpha_b)\Delta(\theta_b^{\mathrm{UN}}) \\
\left|\mathbb{F}_a^{\mathtt{DP}}(\widehat{\theta}_a^{\mathrm{UN}}) - \mathbb{F}_b^{\mathtt{DP}}(\widehat{\theta}_b^{\mathrm{UN}})\right| &= (\alpha_a - \alpha_b)\Delta(\widehat{\theta}_b^{\mathrm{UN}}) + \alpha_a(\mathbb{F}_b^1(\widehat{\theta}_b^{\mathrm{UN}}) - \mathbb{F}_a^1(\widehat{\theta}_a^{\mathrm{UN}})) \\
&\quad + (1-\alpha_a)(\mathbb{F}_b^0(\widehat{\theta}_b^{\mathrm{UN}}) - \mathbb{F}_a^0(\widehat{\theta}_a^{\mathrm{UN}})) \\
&> (\alpha_a - \alpha_b)\Delta(\widehat{\theta}_b^{\mathrm{UN}})
\end{aligned}$$

Since $\theta_b^{\text{UN}} > \widehat{\theta}_b^{\text{UN}} > x_b^*$ (by Theorem 22), $\Delta(\widehat{\theta}_b^{\text{UN}}) > \Delta(\theta_b^{\text{UN}})$.

Therefore, $\left|\mathbb{F}_a^{\text{DP}}(\theta_a^{\text{UN}}) - \mathbb{F}_b^{\text{DP}}(\theta_b^{\text{UN}})\right| < \left|\mathbb{F}_a^{\text{DP}}(\widehat{\theta}_a^{\text{UN}}) - \mathbb{F}_b^{\text{DP}}(\widehat{\theta}_b^{\text{UN}})\right|$.

Next, we show that the disadvantaged group can be flipped under some cost random variable C_a.

Given any $\alpha_a \in (\alpha_b, \delta_u)$, let $(\eta^{\mathcal{C}}(\theta_b^{\text{UN}}), \theta_b^{\text{UN}})$ be a pair of thresholds satisfying fairness $\mathcal{C}$, then if $\Psi_a'(\Delta(\eta^{\mathcal{C}}(\theta_b^{\text{UN}}))) \geq u_-(1-\alpha_a) = \Psi_a'(\Delta(\overline{x_a}))$, we have $\Delta(\eta^{\mathcal{C}}(\theta_b^{\text{UN}})) \geq \Delta(\overline{x_a})$ implying $\eta^{\mathcal{C}}(\theta_b^{\text{UN}}) \leq \overline{x_a}$. Since $\theta_a^{\text{UN}} > \overline{x_a}$, $\eta^{\mathcal{C}}(\theta_b^{\text{UN}}) < \theta_a^{\text{UN}}$ must hold.

Therefore, $\mathbb{F}_b^{\mathcal{C}}(\theta_b^{\text{UN}}) = \mathbb{F}_a^{\mathcal{C}}(\eta^{\mathcal{C}}(\theta_b^{\text{UN}})) < \mathbb{F}_a^{\mathcal{C}}(\theta_a^{\text{UN}})$.

Lastly, we show that cost C_a mentioned above always exists.

Since $\Psi_a'(z) = u_-(1-\alpha_a)(\mathbb{F}_{C_a}(z) + zP_{C_a}(z))$, condition $\Psi_a'(\Delta(\eta^{\mathcal{C}}(\theta_b^{\text{UN}}))) \geq u_-(1-\alpha_a)$ is equivalent to $\mathbb{F}_{C_a}(z) + zP_{C_a}(z) \geq 1$ with $z = \Delta(\eta^{\mathcal{C}}(\theta_b^{\text{UN}}))$, which is attainable. Similarly, the condition in Eqn. (E.2) is equivalent to $\mathbb{F}_{C_a}(z) + zP_{C_a}(z) = c$ for some $c < 1$ with $z = \Delta(\theta_b^{\text{UN}})$, which is also attainable.

E.14 Proof of Proposition 7

Consider the case when $\alpha_a, \alpha_b > \delta_u$. WLOG, let $s := a$, $-s := b$.

1. $\mathcal{C} = \texttt{EqOpt}$: $\mathcal{P}_s^{\texttt{EqOpt}}(x) = f_s^0(1)$

   Because $X|Y = y, S = s, y = \{0,1\}, s = \{a,b\}$ have the same variance σ^2, and $\mu_a^1 - \mu_a^0 < \mu_b^1 - \mu_b^0$, we have $x_s^* = \frac{\mu_s^1 + \mu_s^0}{2}$ and $\mathbb{F}_a^{\texttt{EqOpt}}(x_a^*) > \mathbb{F}_b^{\texttt{EqOpt}}(x_b^*)$.

   When $\alpha_b > \delta_u$, we have $\theta_b^{\text{UN}} < x_b^*$ and $x_b^{\text{UN}} > x_b^*$. As α_b increases, x_b^{UN} and $\mathbb{F}_b^{\texttt{EqOpt}}(x_b^{\text{UN}})$ increase; as $\alpha_b \to \delta_u$, $x_b^{\text{UN}} \to x_b^*$. Therefore, $\exists \omega > \delta_u$ s.t. when $\alpha_b = \omega$, the consequent x_b^{UN} satisfies $\mathbb{F}_a^{\texttt{EqOpt}}(x_a^*) = \mathbb{F}_b^{\texttt{EqOpt}}(x_b^{\text{UN}})$. For any $\alpha_b < \omega$, $\mathbb{F}_a^{\texttt{EqOpt}}(x_a^*) > \mathbb{F}_b^{\texttt{EqOpt}}(x_b^{\text{UN}})$ holds.

2. $\mathcal{C} = \texttt{DP}$: $\mathcal{P}_s^{\text{DP}}(x) = P_{X|S}(x|s) = \alpha_s f_s^0(1) + (1-\alpha_s) f_s^0(x)$.

   Since $\mathbb{F}_{X|Y,S}(x|1,s) < \mathbb{F}_{X|Y,S}(x|0,s), \forall x$, as α_a increases, $\mathbb{F}_a^{\text{DP}}(x_a^*)$ decreases.

   Because $X|Y = y, S = s, y = \{0,1\}, s = \{a,b\}$ have the same variance σ^2, we have $\frac{\mathbb{F}_a^1(x_a^*) - \mathbb{F}_b^1(x_b^*)}{\mathbb{F}_b^0(x_b^*) - \mathbb{F}_a^0(x_a^*)} = 1$. If $\frac{u_+}{u_-} < 1$, $\frac{u_+}{u_-} < \frac{\mathbb{F}_a^1(x_a^*) - \mathbb{F}_b^1(x_b^*)}{\mathbb{F}_b^0(x_b^*) - \mathbb{F}_a^0(x_a^*)}$, which implies that $\delta_u \mathbb{F}_a^1(x_a^*) + (1-\delta_u)\mathbb{F}_a^0(x_a^*) > \delta_u \mathbb{F}_b^1(x_b^*) + (1-\delta_u)\mathbb{F}_b^0(x_b^*)$, i.e., $\mathbb{F}_a^{\text{DP}}(x_a^*) > \mathbb{F}_b^{\text{DP}}(x_b^*)$ when $\alpha_a = \alpha_b = \delta_u$.

   As $\alpha_b \to \delta_u$, $x_b^{\text{UN}} \to x_b^*$. As such, there exist $\omega_1, \omega_2 > \delta_u$ such that $\forall \alpha_b < \omega_1$ and $\forall \alpha_a < \omega_2$, we have $\mathbb{F}_a^{\text{DP}}(x_a^*) > \mathbb{F}_b^{\text{DP}}(x_b^{\text{UN}})$.

The case when $\alpha_a, \alpha_b < \delta_u$ can be proved similarly and is omitted.

E.15 Proof of Proposition 8

WLOG, let $s := a$ and $-s := b$. Let x_s^{UN} be defined s.t. $\Delta_s(x_s^{\text{UN}}) = \Delta_s(\theta_s^{\text{UN}})$ and $x_s^{\text{UN}} \neq \theta_s^{\text{UN}}$ when $\theta_s^{\text{UN}} \neq x_s^*$,

Since $f_a^y(x) = f_b^y(x)$, $x_a^* = x_b^*$ holds. If $U_s(\theta)$ has multiple extreme points, then according to Theorem 22, all extreme points fall between x_s^{UN} and θ_s^{UN}.

Since $\alpha_a > \delta_u > \alpha_b$, $U_a(\theta)$ is increasing over $(-\infty, \theta_a^{\text{UN}})$ and decreasing over $(x_a^{\text{UN}}, +\infty)$, while $U_b(\theta)$ is increasing over $(-\infty, x_b^{\text{UN}})$ and decreasing over $(\theta_b^{\text{UN}}, +\infty)$.

- $\mathcal{C} = \texttt{EqOpt}$

  Since $f_a^y(x) = f_b^y(x)$, $\theta_a^{\texttt{EqOpt}} = \theta_b^{\texttt{EqOpt}}$. To disincentivize under EqOpt fairness, one of the following four possibilities must hold: (1) $\theta_a^{\texttt{EqOpt}} > x_a^{\text{UN}}, \theta_b^{\texttt{EqOpt}} < x_b^{\text{UN}}$ (2) $\theta_a^{\texttt{EqOpt}} < \theta_a^{\text{UN}}, \theta_b^{\texttt{EqOpt}} > \theta_b^{\text{UN}}$ (3) $\theta_a^{\texttt{EqOpt}} < \theta_a^{\text{UN}}, \theta_b^{\texttt{EqOpt}} < x_b^{\text{UN}}$ (4) $\theta_a^{\texttt{EqOpt}} > x_a^{\text{UN}}, \theta_b^{\texttt{EqOpt}} > \theta_b^{\text{UN}}$.

  Note that (3) and (4) never hold.

  Suppose (3) (resp. (4)) holds, then always $\exists(\theta'_a, \theta'_b)$ satisfying EqOpt with $\theta'_a > \theta_a^{\texttt{EqOpt}}, \theta'_b > \theta_b^{\texttt{EqOpt}}$ (resp. $\theta'_a < \theta_a^{\texttt{EqOpt}}, \theta'_b < \theta_b^{\texttt{EqOpt}}$) s.t. (θ'_a, θ'_b) attains a higher utility. In other words, $(\theta_a^{\texttt{EqOpt}}, \theta_b^{\texttt{EqOpt}})$ cannot be optimal fair policies. It concludes that (3) and (4) cannot hold.

  Note that (1) and (2) cannot be satisfied, because $x_b^{\text{UN}} < x_b^* = x_a^* < x_a^{\text{UN}}$, $\theta_b^{\text{UN}} > x_b^* = x_a^* > \theta_a^{\text{UN}}$, and $\theta_a^{\texttt{EqOpt}} = \theta_b^{\texttt{EqOpt}}$ must hold.

  Therefore, none of four cases can be satisfied. EqOpt cannot disincentivize both groups.

- $\mathcal{C} = \texttt{DP}$

  To disincentivize under DP fairness, one of the following four possibilities must hold: (1) $\theta_a^{\texttt{DP}} > x_a^{\text{UN}}, \theta_b^{\texttt{DP}} < x_b^{\text{UN}}$ (2) $\theta_a^{\texttt{DP}} < \theta_a^{\text{UN}}, \theta_b^{\texttt{DP}} > \theta_b^{\text{UN}}$ (3) $\theta_a^{\texttt{DP}} < \theta_a^{\text{UN}}, \theta_b^{\texttt{DP}} < x_b^{\text{UN}}$ (4) $\theta_a^{\texttt{DP}} > x_a^{\text{UN}}, \theta_b^{\texttt{DP}} > \theta_b^{\text{UN}}$.

  Similar as the case when $\mathcal{C} = \texttt{EqOpt}$, (3) and (4) never hold.

  Note that in order to satisfy DP, it is impossible for (2) to hold. Because $\alpha_a > \alpha_b$ and $f_a^y(x) = f_b^y(x)$, $\theta_b^{\texttt{DP}} < \theta_a^{\texttt{DP}}$ must hold under DP. Moreover, $\theta_a^{\text{UN}} < x_a^* = x_b^* < \theta_b^{\text{UN}}$. Therefore, (2) never hold.

  However, (1) is likely to be satisfied.

  When $U_a(\theta)$, $U_b(\theta)$ have unique extreme point.

Re-write x_s^{UN} as a function of α_s: $x_s^{\mathrm{UN}}(\alpha_s)$, take derivative of $\mathbb{F}_b^{\mathrm{DP}}(x_s^{\mathrm{UN}}(\alpha_s))$ w.r.t. α_s, we have

$$\frac{d\mathbb{F}_s^{\mathrm{DP}}(x_s^{\mathrm{UN}}(\alpha_s))}{d\alpha_s} = \underbrace{\mathbb{F}_s^1(x_s^{\mathrm{UN}}(\alpha_s)) - \mathbb{F}_s^0(x_s^{\mathrm{UN}}(\alpha_s))}_{\textbf{term 1}=-\Delta_s(x_s^{\mathrm{UN}}(\alpha_s))} + \underbrace{P_{X|S}(x_s^{\mathrm{UN}}(\alpha_s)|s) \cdot \frac{dx_s^{\mathrm{UN}}(\alpha_s)}{d\alpha_s}}_{\textbf{term 2}}$$

Note that $\lim_{\alpha_a \to 1} \mathbb{F}_a^{\mathrm{UN}}(x_a^{\mathrm{UN}}(\alpha_a)) = \mathbb{F}_a^{\mathrm{UN}}(+\infty) = 1$, $\lim_{\alpha_b \to 0} \mathbb{F}_b^{\mathrm{UN}}(x_b^{\mathrm{UN}}(\alpha_b)) = \mathbb{F}_b^{\mathrm{UN}}(-\infty) = 0$,

$\lim_{\alpha_a \to \delta_u} \mathbb{F}_a^{\mathrm{UN}}(x_a^{\mathrm{UN}}(\alpha_a)) = \delta_u \mathbb{F}_a^1(x_a^*) + (1-\delta_u)\mathbb{F}_a^0(x_a^*)$,

$\lim_{\alpha_b \to \delta_u} \mathbb{F}_b^{\mathrm{UN}}(x_b^{\mathrm{UN}}(\alpha_b)) = \delta_u \mathbb{F}_b^1(x_b^*) + (1-\delta_u)\mathbb{F}_b^0(x_b^*)$.

Since $x_a^* = x_b^*$, $\lim_{\alpha_b \to \delta_u} \mathbb{F}_b^{\mathrm{UN}}(x_b^{\mathrm{UN}}(\alpha_b)) = \lim_{\alpha_a \to \delta_u} \mathbb{F}_a^{\mathrm{UN}}(x_a^{\mathrm{UN}}(\alpha_a))$.

If $\Delta_b(x_b^*) > P_{X|S}(x_b^*|b) \cdot \frac{dx_b^{\mathrm{UN}}(\alpha_b)}{d\alpha_b}\Big|_{\alpha_b=\delta_u}$ (for a special case where $X|Y=y, S=s$ is Gaussian distributed, it can be satisfied if $X|Y=1, S=s$ and $X|Y=0, S=s$ are sufficiently separable), then $\frac{d\mathbb{F}_b^{\mathrm{DP}}(x_b^{\mathrm{UN}}(\alpha_b))}{d\alpha_b}\Big|_{\alpha_b=\delta_u} < 0$, and $\exists \mathcal{I} \subset (0, \delta_u)$ such that $\forall \alpha_b \in \mathcal{I}$, we have $\mathbb{F}_b^{\mathrm{DP}}(x_b^{\mathrm{UN}}(\alpha_b)) > \lim_{\alpha_a \to \delta_u} \mathbb{F}_a^{\mathrm{UN}}(x_a^{\mathrm{UN}}(\alpha_a))$

Therefore, $\exists(\alpha_a, \alpha_b)$ with $\alpha_a \to \delta_u$ and $\alpha_b \in \mathcal{I}$ s.t. $\mathbb{F}_b^{\mathrm{DP}}(x_b^{\mathrm{UN}}(\alpha_b)) > \mathbb{F}_a^{\mathrm{UN}}(x_a^{\mathrm{UN}}(\alpha_a))$.

In this case, if n_a is sufficiently large, we have $\theta_a^{\mathrm{DP}} > x_a^{\mathrm{UN}}$ and $\theta_b^{\mathrm{DP}} < x_b^{\mathrm{UN}}$.

BIBLIOGRAPHY

[1] Target corporation to pay $2.8 million to resolve eeoc discrimination finding. In *U.S. Equal Employment Opportunity Commission*, 2015. https://bit.ly/2RYWn2u.

[2] FTC releases 2019 privacy and data security update. https://bit.ly/3wpzXpT, 2020.

[3] New FTC data shows that the FTC received nearly 1.7 million fraud reports, and FTC lawsuits returned $232 million to consumers in 2019. https://bit.ly/3cDGdm8, 2020.

[4] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM, 2016.

[5] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

[6] Abhay P Aneja and Carlos F Avenancio-León. No credit for time served? incarceration and credit-driven crime cycles. 2019.

[7] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks. ProPublica*, 23, 2016.

[8] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[9] Necdet S Aybat and Garud Iyengar. An alternating direction method with increasing penalty for stable principal component pursuit. *Computational Optimization and Applications*, 61(3):635–668, 2015.

[10] Siddharth Barman and Nidhi Rathi. Fair cake division under monotone likelihood ratios. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 401–437, 2020.

[11] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

[12] Yahav Bechavod, Katrina Ligett, Aaron Roth, Bo Waggoner, and Steven Z. Wu. Equal opportunity in online classification with partial feedback. In *Advances in Neural Information Processing Systems 32*, pages 8972–8982. 2019.

[13] Aurélien Bellet, Rachid Guerraoui, Mahsa Taziki, and Marc Tommasi. *Fast and Differentially Private Algorithms for Decentralized Collaborative Machine Learning*. PhD thesis, INRIA Lille, 2017.

[14] Pascal Bianchi, Walid Hachem, and Franck Iutzeler. A stochastic primal-dual algorithm for distributed asynchronous composite optimization. In *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*, pages 732–736. IEEE, 2014.

[15] Reuben Binns, Ulrik Lyngs, Max Van Kleek, Jun Zhao, Timothy Libert, and Nigel Shadbolt. Third party tracking in the mobile ecosystem. In *Proceedings of the 10th ACM Conference on Web Science*, pages 23–31, 2018.

[16] Avrim Blum, Suriya Gunasekar, Thodoris Lykouris, and Nati Srebro. On preserving non-discrimination when combining expert advice. In *Advances in Neural Information Processing Systems*, pages 8376–8387, 2018.

[17] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

[18] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[19] Mark Braverman and Sumegha Garg. The role of randomness and noise in strategic classification. In *1st Symposium on Foundations of Responsible Computing*, 2020.

[20] Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. *The Journal of Machine Learning Research*, 13(1):2617–2654, 2012.

[21] Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 547–555, 2011.

[22] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

[23] Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 224–232. ACM, 2018.

[24] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

[25] Yatong Chen, Jialu Wang, and Yang Liu. Strategic recourse in linear classification. *arXiv preprint arXiv:2011.00355*, 2020.

[26] Yifang Chen, Alex Cuellar, Haipeng Luo, Jignesh Modi, Heramb Nemlekar, and Stefanos Nikolaidis. Fair contextual multi-armed bandits: Theory and experiments. *arXiv preprint arXiv:1912.08055*, 2019.

[27] Stephen Coate and Glenn C Loury. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, pages 1220–1240, 1993.

[28] Sam Corbett-Davies, Emma Pierson, Avi Feller, and Sharad Goel. A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. In *Washington Post*, 2016. https://wapo.st/3guwxNn.

[29] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

[30] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. https://reut.rs/2RTe4Ak, 2018.

[31] Christos Dimitrakakis, Yang Liu, David Parkes, and Goran Radanovic. Bayesian fairness. In *AAAI*, 2019.

[32] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

[33] DOT. traffic volume counts, nyc opendata. https://bit.ly/2UaUjoV, 2011.

[34] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. http://archive.ics.uci.edu/ml.

[35] Cynthia Dwork. Differential privacy. In *Proceedings of the 33rd international conference on Automata, Languages and Programming-Volume Part II*, pages 1–12. Springer-Verlag, 2006.

[36] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503. Springer, 2006.

[37] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[38] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference of Fairness, Accountability, and Transparency*, 2018.

[39] Danielle Ensign, Frielder Sorelle, Neville Scott, Scheidegger Carlos, and Venkatasubramanian Suresh. Decision making with limited feedback. In *Algorithmic Learning Theory*, pages 359–367, 2018.

[40] EPA. Epa federal test procedure (ftp). https://bit.ly/3zrhkno, 2008.

[41] Liyue Fan and Li Xiong. An adaptive approach to real-time aggregate monitoring with differential privacy. *IEEE Transactions on Knowledge and Data Engineering*, 26(9):2094–2106, 2014.

[42] Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, pages 1–15, 2013.

[43] Ernst Fehr, Lorenz Goette, and Christian Zehnder. A behavioral account of the labor market: The role of fairness concerns. *Annu. Rev. Econ.*, 1(1):355–384, 2009.

[44] Ferdinando Fioretto and Pascal Van Hentenryck. Optstream: releasing time series privately. *Journal of Artificial Intelligence Research*, 65:423–456, 2019.

[45] Pedro A Forero, Alfonso Cano, and Georgios B Giannakis. Consensus-based distributed support vector machines. *Journal of Machine Learning Research*, 11(May):1663–1707, 2010.

[46] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015.

[47] Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. Predictably unequal? the effects of machine learning on credit markets. *The Effects of Machine Learning on Credit Markets*, 2018.

[48] S. Gade and N. H. Vaidya. Private optimization on networks. In *2018 Annual American Control Conference (ACC)*, pages 1402–1409, June 2018.

[49] Christos Ganos, Timo Ogrzal, Alfons Schnitzler, and Alexander Münchau. The pathophysiology of echopraxia/echolalia: relevance to gilles de la tourette syndrome. *Movement Disorders*, 27(10):1222–1229, 2012.

[50] György Gergely and Gergely Csibra. Sylvia's recipe: The role of imitation and pedagogy in the transmission of cultural knowledge. *Roots of human sociality: Culture, cognition, and human interaction*, pages 229–255, 2006.

[51] Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. In *Advances in Neural Information Processing Systems*, pages 2600–2609, 2018.

[52] Kevin Fox Gotham. Race, mortgage lending and loan rejections in a us city. *Sociological Focus*, 31(4):391–405, 1998.

[53] Nika Haghtalab, Nicole Immorlica, Brendan Lucier, and Jack Z. Wang. Maximizing welfare with incentive-aware evaluation mechanisms. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 160–166, 2020.

[54] MT Hale and M Egerstedty. Differentially private cloud-based multi-agent optimization with constraints. In *American Control Conference (ACC), 2015*, pages 1235–1240. IEEE, 2015.

[55] Shuo Han, Ufuk Topcu, and George J Pappas. Differentially private distributed constrained optimization. *IEEE Transactions on Automatic Control*, 62(1):50–64, 2017.

[56] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

[57] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[58] Drew Harwell. The accent gap. https://wapo.st/3goEqDN, 2018.

[59] Drew Harwell. Amazon's alexa and google home show accent bias, with chinese and spanish hardest to understand. 2018. http://bit.ly/2QFA1MR.

[60] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938. PMLR, 2018.

[61] Bingsheng He, Li-Zhi Liao, Deren Han, and Hai Yang. A new inexact alternating directions method for monotone variational inequalities. *Mathematical Programming*, 92(1):103–118, 2002.

[62] Hoda Heidari and Andreas Krause. Preventing disparate treatment in sequential decision making. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2248–2254, 2018.

[63] Hoda Heidari, Vedant Nanda, and Krishna Gummadi. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. In *International Conference on Machine Learning*, pages 2692–2701, 2019.

[64] Tatiana Homonoff, Rourke O'Brien, and Abigail B Sussman. Does knowing your fico score change financial behavior? evidence from a field experiment with student loan borrowers. *Review of Economics and Statistics*, pages 1–45.

[65] Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1389–1398. International World Wide Web Conferences Steering Committee, 2018.

[66] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 259–268, 2019.

[67] Chunan Huang, Xueru Zhang, Rasoul Salehi, Tulga Ersal, and Anna G Stefanopoulou. A robust energy and emissions conscious cruise controller for connected vehicles with privacy considerations. In *2020 American Control Conference (ACC)*, pages 4881–4886. IEEE, 2020.

[68] Zhenqi Huang, Sayan Mitra, and Nitin Vaidya. Differentially private distributed optimization. In *Proceedings of the 2015 International Conference on Distributed Computing and Networking*, page 4. ACM, 2015.

[69] Bradley E Huitema and Joseph W McKean. Autocorrelation estimation and inference with small samples. *Psychological Bulletin*, 110(2):291, 1991.

[70] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1617–1626. JMLR. org, 2017.

[71] IR James. Estimation of the mixing proportion in a mixture of two normal distributions from simple, rapid measurements. *Biometrics*, pages 265–275, 1978.

[72] Andrew M Jones, James Lomas, and Nigel Rice. Applying beta-type size distributions to healthcare cost regressions. *Journal of Applied Econometrics*, 29(4):649–670, 2014.

[73] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Meritocratic fairness for infinite and contextual bandits. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 158–163. ACM, 2018.

[74] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016.

[75] Christopher Jung, Sampath Kannan, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. Fair prediction with endogenous behavior. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 677–678, 2020.

[76] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 63(6):4037–4049, 2017.

[77] Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[78] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[79] Sampath Kannan, Aaron Roth, and Juba Ziani. Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 240–248. ACM, 2019.

[80] Georgios Kellaris and Stavros Papadopoulos. Practical differential privacy via grouping and smoothing. *Proceedings of the VLDB Endowment*, 6(5):301–312, 2013.

[81] Georgios Kellaris, Stavros Papadopoulos, Xiaokui Xiao, and Dimitris Papadias. Differentially private event sequences over infinite streams. *Proceedings of the VLDB Endowment*, 7(12):1155–1166, 2014.

[82] Jonathan Kelner. An algorithmist's toolkit, 2007.

[83] Moein Khajehnejad, Behzad Tabibian, Bernhard Schölkopf, Adish Singla, and Manuel Gomez-Rodriguez. Optimal decision making under strategic behavior. *arXiv preprint arXiv:1905.09239*, 2019.

[84] Mohammad Mahdi Khalili, Mingyan Liu, and Sasha Romanosky. Embracing and controlling risk dependency in cyber-insurance policy underwriting. *Journal of Cybersecurity*, 5(1):tyz010, 2019.

[85] Mohammad Mahdi Khalili, Parinaz Naghizadeh, and Mingyan Liu. Designing cyber insurance policies: Mitigating moral hazard through security pre-screening. In *International Conference on Game Theory for Networks*, pages 63–73. Springer, 2017.

[86] Mohammad Mahdi Khalili, Parinaz Naghizadeh, and Mingyan Liu. Embracing risk dependency in designing cyber-insurance contracts. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 926–933. IEEE, 2017.

[87] Mohammad Mahdi Khalili, Parinaz Naghizadeh, and Mingyan Liu. Designing cyber insurance policies: The role of pre-screening and security interdependence. *IEEE Transactions on Information Forensics and Security*, 13(9):2226–2239, 2018.

[88] Mohammad Mahdi Khalili, Xueru Zhang, Mahed Abroshan, and Somayeh Sojoudi. Improving fairness and privacy in selection problems. In *Proc. 35th AAAI Conference on Artificial Intelligence*, 2021.

[89] Mohammad Mahdi Khalili, Xueru Zhang, and Mingyan Liu. Contract design for purchasing private data using a biased differentially private algorithm. In *Proceedings of the 14th Workshop on the Economics of Networks, Systems and Computation*, pages 4:1–4:6. ACM, 2019.

[90] Mohammad Mahdi Khalili, Xueru Zhang, and Mingyan Liu. Effective premium discrimination for designing cyber insurance policies with rare losses. In *International Conference on Decision and Game Theory for Security*, pages 259–275. Springer, 2019.

[91] Mohammad Mahdi Khalili, Xueru Zhang, and Mingyan Liu. Incentivizing effort in interdependent security games using resource pooling. In *Proceedings of the 14th Workshop on the Economics of Networks, Systems and Computation*, pages 1–6, 2019.

[92] Mohammad Mahdi Khalili, Xueru Zhang, and Mingyan Liu. Public good provision games on networks with resource pooling. In *Network Games, Control, and Optimization*, pages 271–287. Springer, 2019.

[93] Mohammad Mahdi Khalili, Xueru Zhang, and Mingyan Liu. Resource pooling for shared fate: Incentivizing effort in interdependent security games through cross-investments. *IEEE Transactions on Control of Network Systems*, 2020.

[94] Mohammad Mahdi Khalili, Xueru Zhang, and Mingyan Liu. Designing contracts for trading private and heterogeneous data using a biased differentially private algorithm. *IEEE Access*, 9:70732–70745, 2021.

[95] Niki Kilbertus, Manuel Gomez-Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. Improving consequential decision making under imperfect predictions. *arXiv preprint arXiv:1902.02979*, 2019.

[96] Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 825–844, 2019.

[97] Fengjiao Li, Jia Liu, and Bo Ji. Combinatorial sleeping bandits with fairness constraints. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 1702–1710. IEEE, 2019.

[98] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115. IEEE, 2007.

[99] M. Lichman. UCI machine learning repository, 2013.

[100] Qing Ling, Yaohua Liu, Wei Shi, and Zhi Tian. Weighted admm for fast decentralized network optimization. *IEEE Transactions on Signal Processing*, 64(22):5930–5942, 2016.

[101] Qing Ling and Alejandro Ribeiro. Decentralized linearized alternating direction method of multipliers. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 5447–5451. IEEE, 2014.

[102] Qing Ling, Wei Shi, Gang Wu, and Alejandro Ribeiro. Dlm: Decentralized linearized alternating direction method of multipliers. *IEEE Transactions on Signal Processing*, 63(15):4051–4064, 2015.

[103] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158, 2018.

[104] Lydia T Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 381–391, 2020.

[105] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*, 2017.

[106] Ilan Lobel and Asuman Ozdaglar. Distributed subgradient methods for convex optimization over random networks. *IEEE Transactions on Automatic Control*, 56(6):1291–1306, 2011.

[107] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es, 2007.

[108] Sindri Magnússon, Pradeep Chathuranga Weeraddana, Michael G Rabbat, and Carlo Fischione. On the convergence of an alternating direction penalty method for nonconvex problems. In *Signals, Systems and Computers, 2014 48th Asilomar Conference on*, pages 793–797. IEEE, 2014.

[109] Ed McKenzie. Some simple models for discrete variate time series. *JAWRA Journal of the American Water Resources Association*, 21(4):645–650, 1985.

[110] John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6917–6926. PMLR, 13–18 Jul 2020.

[111] Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 230–239, 2019.

[112] Aryan Mokhtari, Wei Shi, Qing Ling, and Alejandro Ribeiro. Decentralized quadratically approximated alternating direction method of multipliers. In *Signal and Information Processing (GlobalSIP), 2015 IEEE Global Conference on*, pages 795–799. IEEE, 2015.

[113] Hussein Mouzannar, Mesrob I. Ohannessian, and Nathan Srebro. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 359–368, 2019.

[114] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning optimal fair policies. *Proceedings of machine learning research*, 97:4674, 2019.

[115] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

[116] Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

[117] Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Y Narahari. Achieving fairness in the stochastic multi-armed bandit problem. *arXiv preprint arXiv:1907.10516*, 2019.

[118] Rohit Kumar Patra and Bodhisattva Sen. Estimation of a two-component mixture model with applications to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):869–893, 2016.

[119] Julia Paxton, Douglas Graham, and Cameron Thraen. Modeling group loan repayment behavior: New insights from burkina faso. *Economic Development and cultural change*, 48(3):639–655, 2000.

[120] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in Neural Information Processing Systems*, 30:5680–5689, 2017.

[121] Vibhor Rastogi and Suman Nath. Differentially private aggregation of distributed time-series with transformation and encryption. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 735–746. ACM, 2010.

[122] US Federal Reserve. Report to the congress on credit scoring and its effects on the availability and affordability of credit. *Board of Governors of the Federal Reserve System*, 2007.

[123] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *stat*, 1050:22, 2017.

[124] Wei Shi, Qing Ling, Kun Yuan, Gang Wu, and Wotao Yin. On the linear convergence of the admm in decentralized consensus optimization. *IEEE Trans. Signal Processing*, 62(7):1750–1761, 2014.

[125] Changkyu Song, Sejong Yoon, and Vladimir Pavlovic. Fast admm algorithm for distributed optimization with adaptive penalty. In *AAAI*, pages 753–759, 2016.

[126] Karthik Sridharan, Shai Shalev-shwartz, and Nathan Srebro. Fast rates for regularized objectives. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1545–1552. 2009.

[127] Marilyn Strathern. 'improving ratings': audit in the british university system. *European review*, 5(3):305–321, 1997.

[128] Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.

[129] Wei Tang, Chien-Ju Ho, and Yang Liu. Fair bandit learning with delayed impact of actions. *arXiv preprint arXiv:2002.10316*, 2020.

[130] I. Vakilinia, J. Xin, M. Li, and L. Guo. Privacy-preserving data aggregation over incomplete data for crowdsensing. In *2016 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, Dec 2016.

[131] Isabel Valera, Adish Singla, and Manuel Gomez Rodriguez. Enhancing the accuracy and fairness of human decision making. In *Advances in Neural Information Processing Systems*, pages 1769–1778, 2018.

[132] Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems*, pages 2722–2731, 2017.

[133] Hao Wang, Zhengquan Xu, Lizhi Xiong, and Tao Wang. Conducting correlated laplace mechanism for differential privacy. In *International Conference on Cloud Computing and Security*, pages 72–85. Springer, 2017.

[134] Ermin Wei and Asuman Ozdaglar. Distributed alternating direction method of multipliers. In *2012 IEEE 51st Annual Conference on Decision and Control (CDC)*, pages 5445–5450. IEEE, 2012.

[135] William WS Wei. Time series analysis. In *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*. 2006.

[136] Christian H Weiß. Monitoring correlated processes with binomial marginals. *Journal of Applied Statistics*, 36(4):399–414, 2009.

[137] Joshua Williams and J Zico Kolter. Dynamic modeling and equilibria in fair decision making. *arXiv preprint arXiv:1911.06837*, 2019.

[138] Yonghui Xiao and Li Xiong. Protecting locations with differential privacy under temporal correlations. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1298–1309. ACM, 2015.

[139] Zheng Xu, Mário AT Figueiredo, and Tom Goldstein. Adaptive admm with spectral penalty parameter selection. *arXiv preprint arXiv:1605.07246*, 2016.

[140] Zheng Xu, Gavin Taylor, Hao Li, Mário AT Figueiredo, Xiaoming Yuan, and Tom Goldstein. Adaptive consensus admm for distributed optimization. In *International Conference on Machine Learning*, pages 3841–3850, 2017.

[141] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

[142] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness constraints: A flexible approach for fair classification. *J. Mach. Learn. Res.*, 20(75):1–42, 2019.

[143] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[144] Chongjie Zhang and Julie A Shah. Fairness in multi-agent sequential decision-making. In *Advances in Neural Information Processing Systems*, pages 2636–2644, 2014.

[145] Chunlei Zhang, Muaz Ahmad, and Yongqiang Wang. Admm based privacy-preserving decentralized optimization. *IEEE Transactions on Information Forensics and Security*, 14(3):565–580, 2019.

[146] Ruiliang Zhang and James Kwok. Asynchronous distributed admm for consensus optimization. In *International Conference on Machine Learning*, pages 1701–1709, 2014.

[147] Tao Zhang and Quanyan Zhu. Dynamic differential privacy for admm-based distributed classification learning. *IEEE Transactions on Information Forensics and Security*, 12(1):172–187, 2016.

[148] Xueru Zhang, Chunan Huang, Mingyan Liu, Anna Stefanopoulou, and Tulga Ersal. Predictive cruise control with private vehicle-to-vehicle communication for improving fuel consumption and emissions. *IEEE Communications Magazine*, 57(10):91–97, 2019.

[149] Xueru Zhang, Mohammad Mahdi Khalili, and Mingyan Liu. Improving the privacy and accuracy of admm-based distributed algorithms. In *International Conference on Machine Learning*, pages 5796–5805. PMLR, 2018.

[150] Xueru Zhang, Mohammad Mahdi Khalili, and Mingyan Liu. Recycled admm: Improve privacy and accuracy with less computation in distributed algorithms. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 959–965. IEEE, 2018.

[151] Xueru Zhang, Mohammad Mahdi Khalili, and Mingyan Liu. Recycled admm: Improving the privacy and accuracy of distributed algorithms. *IEEE Transactions on Information Forensics and Security*, 15:1723–1734, 2019.

[152] Xueru Zhang, Mohammad Mahdi Khalili, and Mingyan Liu. Long-term impacts of fair machine learning. *Ergonomics in Design*, 28(3):7–11, 2020.

[153] Xueru Zhang, Mohammad Mahdi Khalili, Cem Tekin, and Mingyan Liu. Group retention when using machine learning in sequential decision making: the interplay between user dynamics and fairness. In *Advances in Neural Information Processing Systems*, pages 15243–15252, 2019.

[154] Xueru Zhang and Mingyan Liu. Fairness in learning-based sequential decision algorithms: A survey. *Springer Studies in Systems, Decision and Control, Handbook on RL and Control.*, 2020.

[155] Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18457–18469, 2020.