UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Privacy-Preserving and Fairness in Machine Learning

by Tao Zhang

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Tianqing Zhu

University of Technology Sydney
Faculty of Engineering and Information Technology
January 2022

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Tao Zhang declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Computer Science, Faculty of Engineering and Information at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:

SIGNATURE: Signature removed prior to publication.

[Your Name]

DATE: 06th January, 2021

PLACE: Sydney, Australia

ACKNOWLEDGMENTS

First and foremost, I am extremely grateful to my supervisors, Prof. Tianqing Zhu and Prof. Wanlei Zhou, for their invaluable advice, continuous support, and patience during my PhD study. Their immense knowledge and plentiful experience have encouraged me all the time in my research and daily life. In particular, I am really grateful to Prof. Tianqing for giving me a lot of guidance on how to do research, giving me the space to work on the research that I am interested in, and showing support for my internship.

I would also like to thank Dr Dayong Ye for his technical support on my research and for his invitation to coffee chat almost every week. I would like to thank Dr Angela Huo for being my co-supervisor. I would like to thank Prof. Philip Yu, who gave me suggestions for research and helped me polish papers. I want to thank all external collaborators: Ping Xiong, Zahir Tari, and Philip Yu. Thanks for the insightful discussion and valuable advice from them. Additionally, I would like to express gratitude to all staff in the School of Computer Science for helping PhD students solve issues.

I am fortunate to have been a part of our research group - Cyber Privacy and Safety. I also owe plenty of thanks to my research friends who helped me during my research and daily life during my PhD study: Steven Cheng, Sheng Shen, Mengde Han, Tingting Liao, Congcong Zhu, Yuan Zhao, Chi Liu, Zhuowei Wang, Yanbin Liu, and so many others. It is such an honour to work with all of you. Special thanks to Jing Li; he is always willing to discuss research with me, and he has given me a lot of inspiration.

It is important to strike a balance with life outside the hard work of the lab. I would like to thank my tennis buddies: Jack, Neo, Kenyo, Jason, Jianqiao, Alexander, Huipeng Xue and many others. Special thanks to my friend Steven Holley, who let me understand many aspects of Australia. Most importantly, I am grateful for my family's unconditional, unequivocal, and loving support.

Tao Zhang
Sydney, Australia
January 2022

DEDICATION

To myself . . .

Abstract

Machine learning is widely deployed in society, unleashing its magic in a wide range of applications following the progress of big data and computing power. However, society is beginning to realize that machine learning models designed to help human beings in various tasks may also have a negative impact on human beings, especially in terms of privacy and fairness. In terms of privacy, data are increasingly collected from human beings, and when these data are used for machine learning, data privacy might be compromised. In terms of fairness, machine learning, as a useful decision-making tool, is widely used to allocate resources and opportunities for humans. Many studies have shown that decisions made by these models may be biased against certain populations. Machine learning has passed the stage of only considering model performance, and ethical issues have a decisive impact on the use of machine learning. This thesis mainly studies how to design a fair and private machine learning model to foster private and fair machine learning and develops methods broadly covering different aspects of privacy and fairness to enhance the trade-off between fairness, privacy and model accuracy. Specifically, it makes the following contributions.

- We propose a correlation reduction scheme with feature selection - selecting features considering data correlation and utility. The proposed scheme involves five steps to manage the extent of data correlation, preserve privacy, and support accuracy in the model outputs.

- We present a framework of fair semi-supervised learning in the pre-processing phase, including pseudo labeling, re-sampling, and ensemble learning to improve accuracy and decrease discrimination. We also propose a framework of fair semi-supervised learning in the in-processing phase. The objective function includes a loss for both the classifier and label propagation and fairness constraints over labeled and unlabeled data.

- We study the balance between accuracy, privacy and fairness in deep learning by designing two different early stopping criteria to help analysts choose when to stop training a model to achieve their ideal trade-off.
- We investigate how adversarial examples will skew model fairness. We formulate the problem as an optimization problem: maximizing the model bias with the constraint of the number of adversarial examples and the perturbation scale.

Keywords: Machine learning, Differential privacy, Algorithmic fairness

Table of Contents

List of Figures xiv

List of Tables xix

List of Publications xxi

1 Introduction 1

1.1 Private Machine Learning 2

1.2 Fair Machine Learning 3

1.3 Private and Fair Machine Learning 4

1.4 Research Objectives and Challenges 5

1.5 Thesis Outline 7

2 Background 9

2.1 Differential Privacy in Machine Learning 9

2.1.1 Notation 9

2.1.2 Differential Privacy 9

2.1.3 Differentially Private Machine Learning 12

2.2 Fairness in Machine Learning 15

2.2.1 Discrimination Sources 15

2.2.2 Fairness Metrics 17

2.2.3 Discrimination Removal Methods 18

2.3 Interaction between Privacy and Fairness 21

2.3.1 Simultaneous Enforcement of Privacy and Fairness 21

2.3.2 Mutual Impact of Privacy and Fairness 21

3 Correlated Differential Privacy: Feature Selection in Machine Learning 22

3.1 Introduction 22

3.2 Preliminaries 24

3.2.1 Differential Privacy 24

3.2.2 Feature Selection 25

3.3 Example of Traffic Monitoring 25

3.4 The Extent of Data Correlation 27

3.4.1 Correlated Degree 27

3.4.2 Correlated Sensitivity 28

3.5 Correlation Reduction Based on Feature Selection 29

3.5.1 Overview of the Method 29

3.5.2 The Proposed CR-FS Scheme 29

3.5.3 Discussion 33

3.6 Experiments 34

3.6.1 Experimental Setup 34

3.6.2 Experiments for Data Analysis 36

3.6.3 Experiments for Data Publishing 39

3.6.4 Discussion 42

3.7 Summary 42

4 Fairness in Semi-supervised Learning: Unlabeled Data Help to Reduce Discrimination 43

4.1 Introduction 43

4.2 Background 45

4.2.1 Notations 45

4.2.2 Fairness Metrics 46

4.2.3 Discrimination Sources 47

4.2.4 Bias, Variance and Noise 48

4.3 The Proposed Method 48

4.3.1 Overview of the Fairness-enhanced Sampling Framework 48

4.3.2 Where to Sample 49

4.3.3 How to Sample 50

4.3.4 How to Train the Model 51

4.3.5 Discussion 53
4.4 Discrimination Analysis 53
4.5 Experiment 56
4.5.1 Experiments on Real Data 56
4.5.2 Experiments on Synthetic Data 63
4.5.3 Discussion and Summery 66
4.6 Related Work 67
4.6.1 Pre-processing Methods 67
4.6.2 In-processing Methods 67
4.6.3 Post-processing Methods 68
4.6.4 Comparison with Other Work 68
4.7 Summary 69

5 Fairness Constraints in Semi-supervised Learning 70
5.1 Introduction 70
5.2 Preliminaries 72
5.2.1 Notations 72
5.2.2 Graph-base Regularization 73
5.2.3 The Proposed Framework 74
5.2.4 Fair SSL of Logistic Regression 76
5.2.5 A Case of SVM 79
5.2.6 Discussion 80
5.3 Experiment 80
5.3.1 Experimental Setup 81
5.3.2 Experimental Results of Disparate Impact 82
5.3.3 Experimental Results of Disparate Mistreatment 86
5.3.4 Discussion and Summary 91
5.4 Related Work 92
5.4.1 Fair Supervised Learning 92
5.4.2 Fair Unsupervised Learning 93
5.4.3 Comparing with Other Work 93
5.5 Summary 94

6 Balancing Learning Model Privacy, Fairness, and Accuracy with Early Stopping Criteria 95

6.1 Introduction 95
6.2 Background 97
6.2.1 Notation 97
6.2.2 Fairness Metrics 98
6.2.3 Differential Privacy 99
6.2.4 Early Stopping Criteria 101
6.3 Preliminary Studies 101
6.3.1 Preliminary Experiments 101
6.3.2 Results 102
6.4 Early Stopping Criteria Methods 104
6.4.1 Stopping Criterion 1 105
6.4.2 Stopping Criterion 2 106
6.5 Theoretical Analysis 107
6.5.1 The Impact of DP-SGD on Gradient 107
6.5.2 The Impact of DP-SGD on the Training Process 109
6.5.3 The Changes of the Discrimination Level in the Validation Set and Test Set 110
6.6 Experiment 111
6.6.1 Experimental Setup 111
6.6.2 Accuracy and Discrimination Level with SGD and DP-SGD . . . 113
6.6.3 Effect of Hyperparameters 115
6.6.4 The Effect of Stopping Criteria 117
6.6.5 Discrimination Levels in Validation and Test Sets 118
6.6.6 Positive Prediction Rate and True Positive Rate 119
6.6.7 Discussion and Summary 121
6.7 Related work 124
6.7.1 Differentially Private Machine Learning 124
6.7.2 Fair Machine Learning 125
6.7.3 Comparison with Other Work 126
6.8 Summary 126

7 Revisiting Model Fairness via Adversarial Examples 127
7.1 Introduction 127
7.2 Related Work 130
7.3 Vulnerability of Model Fairness to Adversarial Attacks 131
7.3.1 Vulnerability of Individual Fairness 131
7.3.2 Vulnerability of Group Fairness 133
7.4 Method 135
7.4.1 Problem Formulation 135
7.4.2 Adversarial Attack on Individual Fairness 135
7.4.3 Adversarial Attack on Group Fairness 137
7.5 Experiment 139
7.5.1 Experimental Setup 140
7.5.2 Experimental Results 141
7.6 Summary 144

8 Conclusion and Future Work 145
8.1 Conclusion 145
8.2 Future Work 146

Bibliography 147

List of Figures

Figure **Page**

1.1 The interaction between accuracy, privacy and fairness in learning models . 5

3.1 Data correlation for different number of features 36

3.2 Privacy-Accuracy trade-off in SVM for different datasets 37

3.3 Privacy-Accuracy trade-off in LR for different datasets 38

3.4 MAE performance for count queries 40

3.5 MAE performance for mean queries 41

4.1 The three phases of the fairness-enhanced sampling framework: 1) where to sample, 2) how to sample and 3) how to train the model. Step 1 is to generate a new training dataset which consists of the original dataset and the pseudo labeled dataset. Step 2 is to construct multiple fair datasets through re-sampling. Step 3 is to train a model with each of the fair datasets through ensemble learning to produce the final predictions. 49

4.2 The trade-off between accuracy (Red) and discrimination level (Blue). (a) LR in Health dataset; (b) SVM in Health dataset; (c) LR in Bank dataset; (d) SVM in Bank dataset; (e) LR in Adult dataset; (f) SVM in Adult dataset. The X-axis is the sample ratio ρ, which denotes that the percentage of ρ unlabeled data are sampled from the unlabeled dataset and then pseudo labeled for training. 59

4.3 The impact of ensemble learning on the accuracy (Red) and discrimination level (Blue) on (a) LR in Health dataset; (b) SVM in Health dataset; (c) LR in Bank dataset; (d) SVM in Bank dataset; (e) LR in Adult dataset; (f) SVM in Adult dataset. Initially, there is not an obvious link between accuracy and discrimination level. However, as the ensemble size grows, the accuracy and discrimination levels begin to converge. Each point is an average of 50 times. 60

4.4 The impact of sample size on accuracy (Red) and discrimination level (Blue) on (a) LR in Health dataset; (b) SVM in Health dataset; (c) LR in Bank dataset; (d) SVM in Bank dataset; (e) LR in Adult dataset; (f) SVM in Adult dataset. An increasing in the sampling size leads to an increase in accuracy and may help to reduce discrimination level. 62

4.5 Comparison with original scheme (ORI), uniform sampling (US) and preferential sample (PS) with (a) LR in Health dataset; (b) SVM in Health dataset; (c) LR in Bank dataset; (d) SVM in Bank dataset; (e) LR in Adult dataset; (f) SVM in Adult dataset. With the fairness-enhanced sampling method (FS), discrimination decreases without much cost of accuracy or accuracy increases without much cost of discrimination. 64

5.1 The trade-off between accuracy and discrimination in proposed method Semi (Red), FS (Blue), US (Blue cross), PS (Yellow cross) and FES (Green cross) under the fairness metric of disparate impact with LR and SVM in two datasets. As the threshold of covariance c increases, accuracy and discrimination increase. The results demonstrate that our method achieves a better trade-off between accuracy and discrimination than other methods. 83

5.2 The impact of the amount of unlabeled data in the training set on accuracy (Red) and discrimination level (Blue) under the fairness metric of disparate impact with LR and SVM in two datasets. The X-axis is the size of unlabeled dataset; the left y-axis is accuracy, and right y-axis is discrimination level. . 85

5.3 The trade-off between accuracy and discrimination in proposed method Semi (Red), FS (Blue) with LR and SVM in two datasets under the metric of overall misclassification rate. As the threshold of covariance c increases, accuracy and discrimination increase. The results demonstrate that our method using unlabeled data achieves a better trade-off between accuracy and discrimination. 87

5.4 The trade-off between accuracy and discrimination in the proposed method Semi (Red), FS (Blue) with LR and SVM in two datasets under the metric of false negative rate. As the threshold of covariance c increases, accuracy and discrimination increase. 88

5.5 The trade-off between accuracy and discrimination in proposed method Semi (Red), FS (Blue) with LR and SVM in two datasets under the metric of false positive rate. As the threshold of covariance c increases, accuracy and discrimination increase. 90

5.6 The impact of the amount of unlabeled data in the training set on accuracy (Red) and discrimination level (Blue) under the fairness metric of overall mistreatment rate with LR and SVM in two datasets. The X-axis is the size of unlabeled dataset; the left y-axis is accuracy, and the right y-axis is discrimination level. 91

6.1 Challenges in private and fair deep learning: 1) Privacy comes at the cost of accuracy; 2) Fairness comes at the cost of accuracy; 3) Privacy could affect model fairness. 96

6.2 Training with SGD (Blue) and DP-SGD (Red) in the Bank dataset. 103

6.3 Training with SGD (Blue) and DP-SGD (Red) in UTK Face dataset. Figure 3 shows that, with DP-SGD, more variations appear during the training in terms of accuracy and discrimination levels. 104

6.4 Training with SGD (Blue) and DP-SGD (Red) in the Health dataset. 114

6.5 Training with SGD (Blue) and DP-SGD (Red) in the IMDB dataset. Figure 6.5 shows that, with DP-SGD, more variations appear during the training in terms of accuracy and discrimination levels. 115

6.6 Training with SGD (Dark blue) and DP-SGD (other colors; $\sigma \cdot C = 1$ and C varies) in the IMDB dataset. Figure 6.6 shows that clipping bound has an impact on discrimination levels. 116

6.7 Training with SGD (Dark blue) and DP-SGD (other colors; $C = 1$ and σ varies) in the IMDB dataset. Figure 7 shows that the noise scale does not produce a distinctive difference in discrimination levels. 117

6.8 Discrimination level of demographic parity (a) and equal odds (b) on validation set (Blue) and test set (Red) in the Bank dataset. Figure 6.8 shows the changes of discrimination level is similar in validation and test set in the Bank dataset. 119
6.9 Discrimination level of demographic parity (a) and equal odds (b) on validation set (Blue) and test set (Red) in the Health dataset. Figure 6.9 shows the changes of discrimination level is similar in validation and test set in the Health dataset. 120
6.10 Discrimination level of demographic parity (a) and equal odds (b) on validation set (Blue) and test set (Red) in the UTK Face dataset. Figure 6.10 shows the trend of discrimination level is similar in validation and test set. 120
6.11 Discrimination levels on the validation set (Blue) and test set (Red) in the IMDB dataset. Figure 6.11 shows the trend of discrimination levels is similar in both sets. 121
6.12 Training with SGD (Blue) and DP-SGD (Red) in the Health dataset. (a) Positive predication rate in Group 1 (Solid line) and Group 2 (Dotted line) (b) True positive rate in Group 1 (Solid line) and Group 2 (Dotted line). Figure 6.12 shows how PPR and TPR change in two groups in the Health dataset. 122
6.13 Training with SGD (Blue) and DP-SGD (Red) in the Bank dataset. (a) Positive prediction rate in the protected group (solid line) and unprotected group (Dotted line) (b) True positive rate in protected group (solid line) and unprotected group (Dotted line). Figure 6.13 shows how PPR and TPR change in two groups in the Bank dataset. 122
6.14 Training with SGD (Blue) and DP-SGD (Red) in the IMDB dataset. (a) Positive prediction rate in Group 1 (Solid line) and Group 2 (Dotted line) (b) True positive rate in Group 1 (Solid line) and Group 2 (Dotted line). Figure 6.14 shows how PPR and TPR change in two groups in the IMDB dataset. . 123
6.15 Training with SGD (Blue) and DP-SGD (Red) in the UTK Face dataset. (a) Positive predication rate in Group 1 (Solid line) and Group 2 (Dotted line) (b) True positive rate in Group 1 (Solid line) and Group 2 (Dotted line). Figure 6.15 shows how PPR and TPR change for two groups in the UTK Face dataset. 123

7.1 Two examples in the group 1 with negative predicted labels are perturbed into adversarial examples. These original examples are similar to their adversarial examples but are being treated differently by the model. This violates individual fairness, and furthers skews the demographic information in group fairness. 129
7.2 An example of group adversarial bias. 134
7.3 Group adversarial bias and individual group adversarial bias on the German and Bank datasets with increasing perturbation scale ϵ. Results show that adversarial examples in our proposed methods can significantly skew individual fairness and group fairness more than baselines. 143
7.4 The mean of the perturbation norm on the German and Bank datasets with an increasing number of adversarial examples ϵ in Fig. 7.4 (a) and 7.4 (b). The result of the group adversarial bias on the German and Bank datasets with the growing number of adversarial examples in Fig. 7.4 (c) and 7.4 (d). 143

List of Tables

Table **Page**

2.1 Notations . . . 10

3.1 Users' locations at different times . . . 26
3.2 The sum counts of users' locations . . . 26
3.3 Number of features in different stages . . . 35

4.1 Two discriminatory datasets tested on the discriminatory test dataset in ORI method and the proposed fairness-enhanced method (FS) with LR and SVM. We show accuracy (Acc), discrimination level (Dis) and the number of data points of each group in the discriminatory test dataset after classification. . . 65
4.2 Two discriminatory datasets tested on the fair test dataset in ORI method and the proposed fairness-enhanced method (FS) with LR and SVM. We show accuracy (Acc), discrimination level (Dis) and the number of data points of each group in the fair test dataset after classification. . . . 65

5.1 The impact of fairness constraints on different datasets in terms of accuracy (Acc) and discrimination level (Dis) under the fairness metric of disparate impact with LR in the Health dataset. . . . 84
5.2 The impact of fairness constraints on different datasets in terms of accuracy (Acc) and discrimination level (Dis) under the fairness metric of disparate impact with LR in the Titanic dataset. . . . 84
5.3 The impact of fairness constraints on different datasets in terms of accuracy (Acc) and discrimination level (Dis) under the fairness metric of overall misclassification rate with LR in the Bank dataset. . . . 89

5.4 The impact of fairness constraints on different datasets in terms of accuracy (Acc) and discrimination level (Dis) under the fairness metric of overall misclassification rate with LR in the Titanic dataset. 89

6.1 Accuracy, discrimination levels and training epoch for the two early stopping criteria with different parameters on the UTK Face dataset 118

6.2 Accuracy, discrimination level and training epoch for the two early stopping criteria with different parameters on the IMDB dataset 119

7.1 Description of datasets 140

7.2 Results for all datasets and all methods concerning fairness metrics and perturbation metrics 142

List of Publications

Published Papers :

1. **Tao Zhang**, Tianqing Zhu, Kun Gao, and Wanlei Zhou. 2021. "Balancing Privacy, Fairness, and Accuracy with Early Stopping Criteria", in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2021.3129592.

2. **Tao Zhang**, Tianqing Zhu, Jing Li, Mengde Han, Wanlei Zhou and Philip Yu, "Fairness in Semi-supervised Learning: Unlabeled Data Help to Reduce Discrimination," in IEEE Transactions on Knowledge and Data Engineering, doi: 10.1109/TKDE.2020. 3002567.

3. **Tao Zhang**, Tianqing Zhu, Ping Xiong, Huan Huo, Zahir Tari and Wanlei Zhou, "Correlated Differential Privacy: Feature Selection in Machine Learning," in IEEE Transactions on Industrial Informatics, vol. 16, no. 3, pp. 2115-2124, March 2020, doi: 10.1109/TII.2019.2936825.

4. **Tao Zhang**, Dayong Ye, Tianqing Zhu, Tingting Liao, and Wanlei Zhou, 2020. "Evolution of cooperation in malicious social networks with differential privacy mechanisms". Neural Computing and Applications, pp.1-16.

5. **Tao Zhang**, Tianqing Zhu, Renping Liu, and Wanlei Zhou, 2020. "Correlated data in differential privacy: Definition and analysis". Concurrency and Computation: Practice and Experience, p.e6015.

6. Chen, Xin, **Tao Zhang**, Sheng Shen, Tianqing Zhu, and Ping Xiong. "An Optimized Differential Privacy Scheme with Reinforcement Learning in VANET." Computers and Security (2021): 102446.

7. Minghao Wang, Tianqing Zhu, **Tao Zhang**, Jun Zhang, Shui Yu, and Wanlei Zhou, 2020. "Security and privacy in 6G networks: New areas and new challenges". Digital Communications and Networks, 6(3), pp.281-291.

8. Xiuting Gu, Tianqing Zhu, Jie Li, **Tao Zhang**, and Wei Ren. "The Impact of Differential Privacy on Model Fairness in Federated Learning." In International Conference on Network and System Security, pp. 419-430. Springer, Cham, 2020.

SUBMITTED PAPERS :

1. **Tao Zhang**, Tianqing Zhu, Mengde Han, Jing Li, Wanlei Zhou and Philip Yu, 2020. "Fairness Constraints in Semi-supervised Learning". arXiv preprint: 2009.06190.

2. **Tao Zhang**, Tianqing Zhu, Jing Li, and Wanlei Zhou. 2021. "Revisiting Model Fairness via Adversarial Examples", submitted to IEEE Transactions on Neural Networks and Learning Systems.

3. Xin Chen, **Tao Zhang**, Ping Xiong, Sheng Shen. 2021. "Trajectory privacy-preserving with multiple obfuscation over road networks in VANET", submitted to Journal of Information Security and Applications.

Chapter 1

Introduction

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate in predicting results without explicit programming, and ML plays an increasingly important role in our daily lives. The possibility is that you use ML in one way or another without you even knowing it. Due to the speed and efficiency of the decision-making process, many aspects of our daily life decisions are outsourced to machine learning algorithms. Examples include face recognition [35], recommendation systems [39], self-driving cars [68], disease detection [106], loan application [41], and etc.

However, society has begun to realize that machine learning models designed to assist human beings in various tasks could also have a negative impact on human beings, especially in privacy and fairness. Privacy leakage in machine learning happens when information from the training dataset or the model is inferred by adversaries. As data are increasingly collected from human activities, such as social networks and wearable devices, data contain personal information. When the information is used for machine learning, privacy is likely to be leaked by privacy attacks [114]. For example, the genetic information of patients can be extracted from a machine learning model when an adversary only knows a person's name and facial recognition system [53].

Meanwhile, fairness has attracted tremendous attention. More than 20 fairness metrics are defined in machine learning [15, 45, 92, 146]. Basically, discrimination in machine learning refers to when an individual person is treated in a disadvantage way compared with other individuals or groups. The main source of discrimination is data, and training a model with biased data will lead to a discriminated model. In fact, almost all large datasets generated by systems based on learning models are biased. For example, Chouldechova [31] found evidence of racial bias in the recidivism prediction tool where black defendants are more likely to be assessed with high risk than white defendants.

Given the widespread use of machine learning to support decisions over loan allocations, insurance coverage, and many other basic precursors to equity, privacy and fairness in machine learning have become significantly important issues. Thus, how to design machine learning algorithms that keep data private and treat individuals equally is critical. In the following, I will introduce more detailed information in three lines: private machine learning, fair machine learning and private and fair machine learning.

1.1 Private Machine Learning

Over the last decade, the connection between humans and data has never been so inseparable. Meanwhile, the era of big data poses new challenges to human data management, especially in data privacy. Privacy-preserving has been adopted by academia and industry to protect individual privacy when datasets are used for training models. In order to guarantee data privacy, privacy-preserving mechanisms are designed to retain data utility while ensuring that the original information will not be disclosed to other individuals or groups. Popular privacy-preserving mechanisms consist of three types of privacy preservation techniques: cryptographic techniques [3], differential privacy [47], and anonymization techniques [97]. Among these privacy techniques, differential privacy is one of the most promising privacy models to protect data privacy in machine learning. The notion of differential privacy was firstly proposed by Dwork et al. [47], which provides a rigorous mathematical framework for defining and protecting privacy. A common method to achieve differential privacy is to add noise to randomize model outputs. It ensures that the adversary cannot distinguish the participation of the individual in the computation even if the adversary knows some background information.

In terms of the position where random noise is added in machine learning, existing research can be classified into four types: input perturbation, output perturbation, objective perturbation and gradient perturbation. Input perturbation means that individual data are randomly perturbed to some extent before they are handed over to the model for learning or analysis to prevent the model from acquiring real data [54]. Output perturbation is a method of adding noise to the optimal parameters obtained from empirical risk minimization [24, 108], and objective perturbation is a method of adding noise to the objective function [25, 56]. Later, differentially private stochastic gradient (DPSGD) was proposed. In DPSGD, noise is added in the process of solving the optimal model parameters using the gradient descent method, ensuring that the entire process meets differential privacy [1, 125, 144].

1.2 Fair Machine Learning

Machine learning is now in wide use as a decision-making tool in many areas, such as job employment, risk assessment, loan approvals and many other basic precursors to equity. However, the popularity of machine learning has raised concerns about whether the decision-making algorithms make are fair to all individuals. Obermeyer et al. found prejudice in health care systems where black patients assigned the same level of risk by the algorithm are sicker than white patients [105]. Recent findings show that unfair machine learning algorithms will affect legal justice, healthcare, and other aspects of human beings. As we move towards a world of machine-assisted predictions for human beings, the fairness of machine learning has become a very cardinal issue. In the future, our ability to design machine learning algorithms that treat all groups equally may be one of the most influential factors in allocating resources and opportunities. As the influence and scope of these risk assessments increase, academics, policymakers, and journalists have raised concerns that the statistical models from which they are derived might inadvertently encode human biases.

Over the past few years, many research have been devoted to designing fairness metrics, such as statistical fairness [15, 31, 146], individual fairness [45, 72, 92] and causal fairness [78, 83]. These approaches and algorithms can be roughly divided into three categories: pre-processing methods, in-processing methods and post-processing methods. Pre-processing methods adjust data distribution [15, 74] or learn new fair representations [94, 124, 147], to relieve some of the tension between accuracy and fairness. In-processing methods add constraints or regularizers to restrict the correlation between labels and sensitive/protected attributes, i.e., traits that can be targeted for discrimination [75, 81, 146]. Post-process methods calibrate training results [60]. These studies mainly focus on addressing the two most crucial fundamental issues in machine learning fairness: how to formalize the concept of fairness in the context of machine learning tasks, and how to design effective algorithms to achieve an ideal compromise between accuracy and fairness.

1.3 Private and Fair Machine Learning

Privacy and fairness are two social concerns in machine learning, and they have a connection with each other. In some situations, privacy, fairness and accuracy are considered at the same time when implementing machine learning models. For example, in recidivism prediction applications, demographic groups (such as black defendants and white defendants) should undergo similar processing, that is, similar prediction accuracy. Meanwhile, participating in training data means that the individual might have committed a crime, which is very sensitive and needs to be kept confidential. Therefore, it is important to study the interaction between privacy, fairness and privacy.

Two perspectives are studied in the interaction between privacy and fairness: 1) simultaneous enforcement of privacy and fairness, 2) mutual impact of privacy and fairness. In the first perspective, the goal is to guarantee privacy and fairness at the same time on machine learning tasks. For example, [40, 139] proposed differentially private fair algorithms in logistic regression. [67] studied fair learning under the constraint of differential privacy in the equalized odds. In the second perspective, it is found that implementing privacy mechanisms, such as differential privacy will have a disparate impact on groups or implementing fairness methods is likely to lead to disparate privacy leakage between groups. For example, Chang et al. studied privacy risks of group fairness through the lens of membership inference attacks [22]. The impact of differential privacy is studied on fair and equitable decision-making [113].

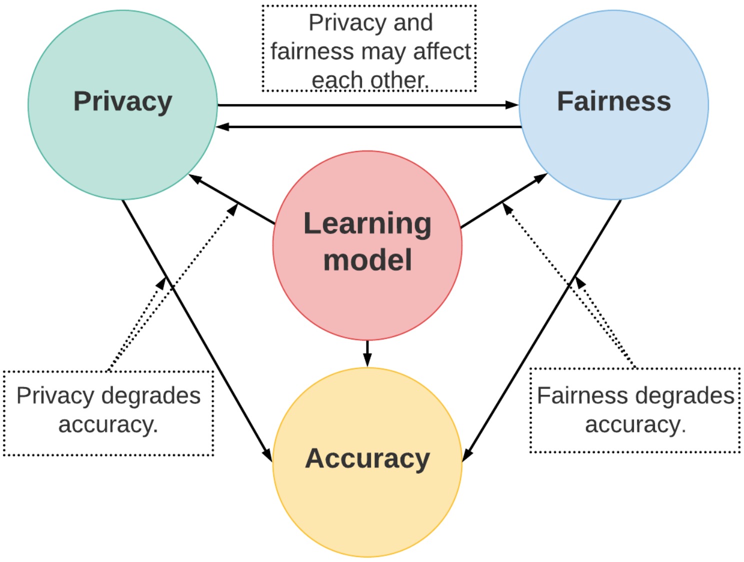

Privacy and fairness may affect each other.
Privacy
Fairness
Learning model
Accuracy
Privacy degrades accuracy.
Fairness degrades accuracy.

Figure 1.1: The interaction between accuracy, privacy and fairness in learning models

1.4 Research Objectives and Challenges

In this section, we first introduce research objectives in private and fair machine learning, and then present research objectives and challenges. As shown in Figure 1.1, model accuracy, privacy and fairness interact with each other. Privacy and fairness are usually at the cost of model accuracy, and meanwhile implementing privacy/fairness methods may affect model fairness/privacy. This thesis will focus on enhancing the trade-off in private/fair learning and analysing the interactions between the three. Detailed objectives and challenges are listed in the following.

- **Objective 1: Privacy leakage in correlated training sets.** Our first research objective is to study privacy leakage in correlated training sets. Data collected from real-world is likely to have correlations, such as temporal correlation and social correlation. Existing differentially private machine learning algorithms have not considered the impact of data correlation in the training set, which may lead to more privacy leakage than expected in real-world applications.

  Challenges. Correlated datasets usually perform a large sensitivity when using differential privacy, which leads to a shape decrease in data utility. How to decrease a large amount of noise incurred by differential privacy in correlated training sets, so as to achieve an ideal trade-off between privacy and data utility is very challenging.

- **Objective 2: Fairness in semi-supervised learning.** In real-world machine learning tasks, data size is an important factor in deciding the model performance. Labeling data takes time and effort, and therefore how training the model with a combination of labeled and unlabeled data is a vital area of development. Like the other learning settings, achieving the balance between accuracy and fairness is a key issue. According to recent studies, increasing the size of the training set can create a better trade-off. This finding sparked an idea over whether the trade-off might be improved via unlabeled data. Unlabeled data are easy to use and, if they could be used as training data, we may be able to make a better trade-off between fairness and accuracy via semi-supervised learning.

  Challenges. To achieve this goal, two challenges are ahead of us: 1) how to achieve fair learning from both labeled and unlabeled data; and 2) how to give labels for unlabeled data to ensure that the learning is towards a fair direction.

- **Objective 3: Simultaneous enforcement of privacy and fairness in deep learning.** Privacy and fairness are two social concerns when applying machine learning models into practice. Implementing privacy methods in the training model will have an impact on model fairness, and vice versa. In some situations, privacy, fairness and accuracy are considered at the same time when implementing machine learning tasks, such as health care systems. How to enforce privacy and fairness simultaneously is a significant issue to be addressed.

  Challenges. Unfortunately, privacy and fairness are usually at the cost of model accuracy. Meanwhile, deep learning is much more complex than traditional machine learning models when dealing with privacy and fairness issues. Challenges include how to define what a good balance between the three is; how to achieve efficient and effective means of fine-tuning the balance between this three.

- **Objective 4: Fairness attack via adversarial examples.** Our last objective is to study the vulnerability of fairness via adversarial examples so that we can have a better understanding of the connection between fairness and adversarial attacks. Existing literature has largely ignored fairness robustness. The fairness of classifiers is often evaluated on sampled datasets, and can be unreliable for various reasons, including biased examples, and missing and/or noisy attributes. Evaluating model fairness is the key to determining the effectiveness of bias removal methods and improving model fairness in dynamic learning systems.

Challenges. This is challenging to define individual adversarial bias and group adversarial bias; explain the vulnerability of individual fairness and group fairness to adversarial attacks; and maximize model bias with a fixed perturbation scale on limited adversarial examples.

1.5 Thesis Outline

This section aims to build the structure of the thesis. In terms of three research problems: correlated differential privacy, fairness in semi-supervised learning and interaction between privacy and fairness in machine learning, the content of each chapter is organised as follows:

- Chapter 2 presents a literature survey on the development of differentially private machine learning, fair machine learning, and private and fair machine learning, including notations, relevant concepts and emerging techniques in this area.

- Chapter 3 addresses the privacy issue for correlated datasets. We propose a correlation reduction scheme with differentially private feature selection considering the issue of privacy loss when data have correlations in machine learning tasks. The proposed scheme involves five steps with the goal of managing the extent of data correlation, preserving privacy, and supporting accuracy in the prediction results.

- Chapter 4 studies how labeled data help to reduce discrimination. We present a framework of fair semi-supervised learning in the pre-processing phase, including pseudo labeling to predict labels for unlabeled data, a re-sampling method to obtain multiple fair datasets and lastly, ensemble learning to improve accuracy and decrease discrimination.

- Chapter 5 focuses on fairness in semi-supervised learning. We develop a framework for fair semi-supervised learning, which is formulated as an optimization problem. This includes classification loss to optimize accuracy, label propagation loss to optimize unlabled data prediction, and fairness constraints over labeled and unlabeled data to optimize the fairness level.

- Chapter 6 studies the balance between accuracy, privacy and fairness in deep learning. We conduct a series of analyses, both theoretical and empirical, on the impacts of implementing DP-SGD in deep neural network models through gradient clipping and noise addition. Based on observations, we designed two different early stopping criteria to help analysts choose the optimal epoch at which to stop training a model so as to achieve their ideal trade-off. Extensive experiments show that our methods can achieve an ideal balance between accuracy, fairness and privacy

- Chapter 7 investigates how adversarial examples will skew model fairness. we study the vulnerability of individual fairness and group fairness to adversarial attacks. We further propose a general adversarial fairness attack framework capable of twisting model bias through a small subset of adversarial examples. We formulate this problem as an optimization problem: maximizing the model bias with the constraint of the number of adversarial examples and the perturbation scale.

- Chapter 8 is the conclusion of the thesis, and gives some possible suggestions and extensions for future research.

To maintain readability, each chapter is organized in a self-contained format. Some contents, such as definitions and related work are introduced in related chapters.

CHAPTER 2

Background

2.1 Differential Privacy in Machine Learning

2.1.1 Notation

Let D be a dataset with N data examples $D = \{(x_i, y_i)\}_{i=1}^{N}$, where $x \sim X$ contains attribute information and $y \sim Y$ denotes the label. Two datasets D and D' are referred to as neighboring datasets when they differ in one example. Consider a function $f : X \rightarrow Y$, which maps training examples x to discrete labels $y \in Y$. Differenial privacy (DP) provides a randomization mechanism $\mathscr{M}$ that can perturb the output of the function f, and the noisy output is denoted as $\hat{Y}$. Table 2.1 lists the notations used in this chapter.

2.1.2 Differential Privacy

Differential Privacy is a privacy model which ensures that changing one record in the dataset will not affect too much of the output. The model is achieved by differential mechanisms which calibrating some noise to the output. DP is a widely used privacy model in machine learning to bound the leakage about the presence of specific data point [25].

Table 2.1: Notations

Notations	**Meanings**
D	dataset
D'	neighbouring dataset
N	the number of data examples
x	attributes
z	sensitive attribute
y	labels
$\hat{y}$	perturbed label
$\mathscr{M}$	randomization mechanism
f	function(or model)
q	query
ϵ	privacy budget
δ	broken probability
Ω	outcome set
Δq	sensitivity of a query
ϕ	output of exponential mechanism
s	score function
σ	noise scale
w	model parameter
$\hat{w}$	perturbed model parameter
$\mathscr{L}$	loss function
r	learning rate
T	the number of training epochs
l	the number of iterations
b	batch size
n	noise value
g	gradient
$\hat{g}$	perturbed gradient
C	clipping bound
d	distance metric
M	distance function
C	clipping bound
γ	the probability of positive predictions
Γ	discrimination level

Definition 2.1. **(ϵ, δ-Differential privacy)** [47] Given neighboring datasets D and D' that differ in one data point, an algorithm $\mathscr{M}$ satisfies (ϵ, δ)-differential privacy for any possible outcome Ω

$$Pr[\mathscr{M}(D) \in \Omega] \leq \exp(\epsilon) \cdot Pr[\mathscr{M}(D') \in \Omega] + \delta, \tag{2.1}$$

where ϵ is privacy budget which determines privacy level, and δ is a broken probability. The lower ϵ and δ represent the higher privacy level.

Definition 2.2. **(Sensitivity)** [47] For a query $q : D \to R$, and neighboring datasets D and D', the sensitivity of q is defined as

$$\Delta q = \max_{D,D'} ||q(D) - q(D')||. \tag{2.2}$$

where $||\cdot||$ denotes L_1 or L_2 norm. Sensitivity measures the maximal difference between neighboring datasets.

DP mechanisms are referred to as the mechanisms that can generate randomized noise to satisfy the requirement in DP. Here, we introduce the most popular mechanisms in differential privacy: laplace mechanism, exponential mechanism and Gaussian mechanism.

Definition 2.3. **(Laplace mechanism)** [48] For any query q: $D \to R$ over a dataset D, the following mechanism provides ϵ-differential privacy if

$$\mathscr{M}(D) = q(D) + Laplace(\Delta q / \epsilon) \tag{2.3}$$

The Laplace noise is denoted as $Laplace(\cdot)$ and is drawn from a Laplace distribution.

Definition 2.4. **(Exponential mechanism)** [48] Given a score function $s(D, q(D))$ of a dataset D, the exponential mechanism $\mathscr{M}$ satisfies ϵ-differential privacy if

$$\mathscr{M}(D) = \left(return\ \phi \propto exp(\frac{\epsilon s(D, q(D))}{2\Delta q}) \right) \tag{2.4}$$

where score function $s(D, q(D))$ is used to evaluate the quality of an output $q(D)$ and Δq is the sensitivity.

Definition 2.5. (**Gaussian mechanism**) [48] Gaussian mechanism adds zero-mean Gaussian noise with variance $\Delta q^2\sigma^2$ in each coordinate of the output $q(D)$

$$\mathcal{M}(D) = q(D) + \mathcal{N}(0, \Delta f^2\sigma^2), \tag{2.5}$$

where σ denotes the noise scale. It satisfies (ϵ,δ)-differential privacy if $\epsilon \in [0,1]$, $\delta \geq c\Delta q/\epsilon$, and $c^2 \geq 2ln(1.25/\delta)$.

DP provides an elegant combination properties, which makes more complex algorithms and data analysis possible by combining multiple differentiated private blocks.

Theorem 2.1. ***(Sequential composition)*** *[48] Suppose that a set of privacy mechanisms $\mathcal{M}$={$\mathcal{M}_1,...,\mathcal{M}_m$}, gives ϵ_i differential privacy ($i = 1,2...,m$), and these mechanisms are sequentially performed on a dataset. $\mathcal{M}$ will provides $(\sum_i \epsilon_i)$-differential privacy for this dataset.*

Theorem 2.2. ***(Parallel composition)*** *[48] Suppose that a set of privacy mechanism $\mathcal{M}$={$\mathcal{M}_1,...,\mathcal{M}_m$}, gives ϵ_i differential privacy ($i = 1,2...,m$), and these mechanisms are performed on the disjoint subsets of an entire dataset. $\mathcal{M}$ will provide max(ϵ_i)-differential privacy for this dataset.*

2.1.3 Differentially Private Machine Learning

2.1.3.1 Basic Knowledge of Machine Learning

According to the different applications of data processing and analysis, machine learning models can be divided into traditional machine learning methods based on statistical learning theory such as linear regression, logistic regression (LR) and support vector machine (SVM), and various neural networks models. For most cases, empirical risk minimization (ERM) is the most commonly used model learning strategy. The basic idea is to search the optimal model parameters w that minimize empirical risk in the whole parameter domain to find the best model f_w mapping between every data examples (X,Y). The definition is as follows

Definition 2.6. (Empirical risk minimization) Given a model f_w, dataset D and loss function $\mathcal{L}$, the empirical risk loss is defined as

$$J_D(f_w) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(f_w(x), y) + \lambda R(w). \tag{2.6}$$

where the loss function $\mathscr{L}(f_w)$ is used to measure the difference between the label y and the classifier prediction $f_w(x)$; a regularizer term $R(w)$ is ususally added to the loss $J(f_w)$ to avoid overfiting from the training data D; λ is a parameter to control the trade-off between classification loss and regularization loss. The optimized model parameters w are solved by minimizing empirical loss via gradient descent. Stochastic gradient descent (SGD) is a widely used method to update the model parameters by computing average gradient on a batch of data examples iteratively. In each iteration l, a batch of data points b are sampled from D and the model parameters are updated by the following rule

$$w^{l+1} = w^l - r\nabla w^l, \tag{2.7}$$

where r is the learning rate and ∇w^l is the gradient of the average loss over a batch of data points. Given a number of T training epochs, the number of iterations is $l = \frac{TN}{b}$.

Due to the simple structure of traditional machine learning models, when designing the loss function $\mathscr{L}$, it will be made as a convex function as much as possible in order to obtain an optimal solution. Deep learning models introduce a large number of non-linear factors, and the objective function is often a non-convex function, so it is easy to fall into a local optima solution. In addition, deep learning also has problems such as a large number of parameters, a large number of iterations, and slow convergence. Therefore, the privacy-preserving methods of the above two types of models are different. In the following, we introduce existing DP methods for these two types of models.

2.1.3.2 Differentially Private Machine Learning Methods

Many works have been studied on differentially private machine learning. In terms of the position where random noise is added in the training process, we classify existing research into four types: 1) input perturbation, 2) output perturbation, 3) objective perturbation and 4) gradient perturbation. The first three methods are used for traditional machine learning models, and the fourth method is mainly for neural network models.

Input Perturbation Input perturbation means that individual data are randomly perturbed to some extent before they are handed over to the model for learning or analysis to prevent the model from acquiring real data [54]. Two situations are considered: 1) Global privacy, that is, personal data is collected centrally first, and then the collector disturbs the sensitive dataset when publishing data; 2) Local privacy means that individuals first disturb the data at the local and then send it to the collector. The former has proved to have great limitations in early studies [6]. In the latter, because the global data distribution is unknown to users, the perturbation mechanism based on global sensitivity is no longer applicable. Thus, local differential privacy (LDP) is proposed for different data types and various data mining tasks [142].

Output Perturbation The second method is output perturbation, which refers to adding noise directly to the optimal parameters obtained by minimizing empirical risk, as shown in the following

$$\hat{w} = \underset{w}{\operatorname{argmin}} J(f_w) + n \tag{2.8}$$

where n denotes noise generated by DP mechanisms. Output perturbation directly adds noise to the real output of the algorithm, so it is the most intuitive perturbation method, and the noise level depends on the sensitivity of $\underset{w}{\operatorname{argmin}} J(w)$ and privacy budget. When the objective function $J(w)$ of empirical risk minimization satisfies continuous differentiable and convex function, $\underset{w}{\operatorname{argmin}} J(w)$ can be obtained with the proof that the output perturbation method satisfies the requirement of DP [24, 108]. However, when the objective function is non-convex, this method is not applicable.

Objective Perturbation Objective perturbation is to minimize the empirical risk by introducing random noise into the objective function expression, ensuring that the optimization process meets difference privacy [25, 56]. The objective function with noise is as follows

$$\hat{\mathscr{L}} = J(f_w) + n^T w \tag{2.9}$$

The value of variable n is generated from DP mechanisms, and the noise level only depends on privacy budget. However, objective perturbation also requires continuous differentiable and convex function so that DP is satisfied.

To solve this issue, a method is proposed to approximately computing objective function by Taylor expansions. In the method, Laplace noise is added to each coefficient. Although this method is successfully applied to the logistic regression model, it is difficult to expand this method to a more general model because the approximate polynomial method is only for a specific objective function.

Gradient Perturbation Gradient perturbation refers to introducing random noise in the process of solving the optimal model parameters by gradient descent method, and ensuring that the whole process meets difference privacy. [126] firstly derived differentially private stochastic gradient descent mechanisms and test them empirically in logistic regression. In order to ensure the computational efficiency of the algorithm, stochastic gradient descent is often used in practical applications, and the basic implementation of gradient perturbation is in the following

$$w^{l+1} = w^l - r(\Delta w^l + n) \tag{2.10}$$

where n is usually generated from Gaussian mechanism. The most popular gradient perturbation method is proposed by Abadi et al [1]. The method mainly includes two steps: gradient clipping and noise addition. The first step is that a clipping bound C is used on the l_2 norm of the gradient updates $g^l(x_i)$ with the batch size b. The second step aggregates clipped gradients $\bar{g}^l$, and adds Gaussian noise $\mathcal{N}(0, \sigma^2 C^2)$ to the aggregation. Since each update of $\hat{g}^l$ is differentially private, the final model parameters are also differentially private in terms of the composition property of differential privacy. In this method, differential privacy noise is added to the iterative update of gradients and ensuring that the entire process meets differential privacy [1, 144].

2.2 Fairness in Machine Learning

2.2.1 Discrimination Sources

Discrimination may exist in many forms, some of which may lead to the unfairness of different downstream learning tasks. In [100], the authors discuss the sources of bias in machine learning, different sources that may affect AI applications, and description to inspire future solutions to each of discrimination sources. Here, we introduce some of the most general discrimination sources.

Historical discrimination Historical discrimination occurs when there is a discrepancy between the world itself and the values or goals in the model to be encoded and propagated. It can stem from cultural stereotypes among people, such as social class, race, nationality and gender. For example, men are considered to have better leadership than women, and the result is that men are more likely to get promotion than women.

Measurement Discrimination Measurement discrimination comes from the way we choose, utilize, and measure specific features. The selected set of features and labels may miss important factors, or bring in group or input-related noise that causes different performance. For example, a minority community are given more police force, and thus a higher arrest rate occurs in this community. However, it is unfair to say that people from this community is more dangerous because there is a difference in how the arrest rate is measured.

Representation Discrimination Representation discrimination occurs when the data used to train the algorithm does not accurately represent the problem space. As a consequence, the model generalizes to fit the majority groups much than minority groups. For example, face recognition tasks show a better performance in the white than the black because the training set contains more white people images than black peoples images.

Aggregation Discrimination Aggregation discrimination can arise during model construction when different populations are improperly grouped together. In many applications, the groups of interest are heterogeneous, so a single model is unlikely to fit all subgroups. For example, diabetes patients have quite different symptoms across gender and race, and these factors will lead to different meanings and importance. When a clinical aid model is trained on the data that is monitored of diabetes in complicated ways across gender and race, a single model is unlikely to be the best suited for all people.

Evaluation Discrimination Evaluation discrimination occurs during the model iteration and evaluation. This can happen when a test or external benchmark unequally represents each group in the population. Evaluation discrimination may also occur due to the use of performance metrics that are not appropriate for the way the model is used.

2.2.2 Fairness Metrics

More than 20 fairness metrics have been proposed to define what is fairness in machine learning [19]. Here, we mainly focus on individual fairness and group fairness.

Individual Fairness Individual fairness is formalized by viewing machine learning models as maps between input and output metric spaces and defining individual fairness as Lipschitz continuity of machine learning models. The definition is given as follows.

Definition 2.7. (**Individual Fairness**) [45] A model $f : X \to Y$ is individually fair if for all $x_i, x_j \in X$, we have $M(f(x_i), f(x_j)) \leq d(x_i, x_j)$.

where M is a distance function that measures the difference in the probabilities. Individual fair models require that any two data examples x_i, x_j that are at distance $d(x_i, x_j)$, and map to distributions $f(x_i)$ and $f(x_j)$, respectively, such that the statistical distance between $f(x_i)$ and $f(x_j)$ is $d(x_i, x_j)$ at most. The distance metric d on the input space depends on the machine learning tasks because it encodes the intuition as to which users are similar.

Group Fairness In group fairness, we first need to know what sensitive attributes are. Research focuses on different intuitive concepts of "unfair decisions", which are usually considered to be based on a certain personal attribute, such as gender, race, age, sexual orientation, political and religious orientation, which can influence people's decisions in different ways. These attributes are considered as protected or sensitive attributes.

Data examples $x_i \sim X$ contain the information of k unprotected attributes, and the protected attribute z (also called sensitive attribute). When considering a binary protected attribute, D is divided into the advantaged group and the disadvantaged group. The advantaged group is referred to as the group that has better performance in terms of a fairness metric, and vice versa. For example, if the protected attribute is 'Sex' and the learning task is face recognition, the dataset can be divided into black and white groups. Because the white group has better performance, the white group is the advantaged group and the black group is the disadvantaged group. The training goal is to learn a mapping f_w over the dataset D to well approximately mapping between input X and output Y in a fair way.

Definition 2.8. (Demographic parity) [15] A classification model satisfies demographic parity if

$$Pr(\hat{y}=1|z=1)=Pr(\hat{y}=1|z=0), \tag{2.11}$$

where $\hat{y}$ is the predicted label.

Definition 2.9. (Equal opportunity) [60] Equal opportunity requires that the true positive rates are equal in different groups, which is presented as

$$Pr(\hat{y}=1|z=1,y=1)=Pr(\hat{y}=1|z=0,y=1). \tag{2.12}$$

Definition 2.10. (Equal odds) [60] Equal odds requires that the predicted result of a classifier is independent of the sensitive attribute on the condition of true positive class and false positive rate, which is given as

$$Pr(\hat{y}=1|z=1,y=0)=Pr(\hat{y}=1|z=0,y=0), \tag{2.13}$$

$$Pr(\hat{y}=1|z=1,y=1)=Pr(\hat{y}=1|z=0,y=1). \tag{2.14}$$

Definition 2.11. (Discrimination level) Let $\gamma_z(D,f_w)$ denote the probability of positive predictions of group z on a model f_w training with a dataset D in terms of a fairness metric. The discrimination level $\Gamma(D,f_w)$ on a model f_w training with a dataset D is measured by the difference between groups:

$$\Gamma(D,f)=|\gamma_0(D,f_w)-\gamma_1(D,f_w)|. \tag{2.15}$$

Take demographic parity as an example, we have $\gamma_1(D,f_w)=Pr(\hat{y}=1|z=1)$, and the discrimination level is $\Gamma(D,f_w)=|Pr(\hat{y}=1|z=1)-Pr(\hat{y}=1|z=0)|$.

2.2.3 Discrimination Removal Methods

Discrimination removal methods are generally classified into three streams.

2.2.3.1 Fair Supervised Learning

Methods for fair supervised learning include pre-processing, in-processing and post-processing methods.

Pre-processing Pre-processing methods recognize that the bias is often the data itself, where the distribution of sensitive attributes is biased, discriminatory, or unbalanced. In pre-processing, discrimination is eliminated by guiding the distribution of training data towards a fairer direction [73] or by transforming the training data into a new space [16, 52, 124, 147, 151]. Commonly used methods in pre-processing include re-sampling [73, 130], re-weighting [15, 73], adversarial learning [52, 94], causal methods [78, 83] and transformation [59, 147]. The main advantage of pre-processing methods is that it does not require changes to the machine learning algorithm, so it is very simple to use.

In-processing In-processing methods recognize that the learning model is often affected by important features and other data distribution effects, or try to find a balance between multiple model goals. In-processing techniques include adversarial learning [49, 148], constraint optimization [4, 43, 57, 145, 146], and regularization [5, 66, 75]. For example, [57, 145, 146] designed the convex fairness constraint, called decision boundary covariance to achieve fair classification for classifiers. This category is more flexible for optimizing different fair constraints, and solutions using this method are considered to be the most robust. In addition, these methods have shown good results in terms of accuracy and fairness.

Post-processing The post-processing methods recognize that the output of models may be unfair to one or more of the protected attributes for different groups. Post-processing techniques include calibration [90, 111], threshold [60, 79, 91]. For example, In [60], a learned classifier is modified to adjust the decisions to be non-discriminatory for different groups by adjusting the threshold. Post-processing does not need changes in the classifier and only needs to access prediction and sensitive attribute information, without access to actual algorithms and models. This makes them suitable for black-box scenarios, which do not include the exposure of the entire machine learning pipeline.

2.2.3.2 Fair Unsupervised Learning

Chierichetti et al. [30] was the first to study fairness in clustering. Their solution, under both k-center and k-median objectives, required every group to be (approximately) equally represented in each cluster. Many subsequent works have since been undertaken on the subject of fair clustering. Among these, Rosner et al. [116] extended fair clustering to more than two groups. Chen et al. [118] consider the fair k-means problem in the streaming model, define fair coresets and show how to compute them in a streaming setting, resulting in a significant reduction in the input size. Bera et al. [11] presented a more generalized approach to fair clustering, providing a tunable notion of fairness in clustering. Chen et al. [29] proposed a notion of proportionally fair clustering where all possible large-scale groups have the right to choose their centers according to the concept of proportional fair clustering. Kleindessner et al. [80] studied a version of constrained spectral clustering incorporating the fairness constraints.

2.2.3.3 Fair Semi-supervised Learning

Existing fair learning methods mainly focus on supervised and unsupervised learning, and cannot be directly applied to semi-supervised Learning (SSL). As far as we know, only [32, 104] has explored fairness in SSL. Chzhen et al. [32] studied Bayes classifier under the fairness metric of equal opportunity, where labeled data is used to learn the output conditional probability, and unlabeled data is used to calibrate the threshold in the post-processing phase. However, unlabeled data is not fully used to eliminate discrimination, and the proposed method only applies to equal opportunity. In [104], the proposed method is built on neural networks for SSL in the in-processing phase, where unlabeled data is marked labels with pseudo labeling. In this thesis, we proposed a pre-processing framework that includes pseudo labeling, re-sampling and ensemble learning to remove representation discrimination in semi-supervised learning [149]. We also propose solutions for a margin-based classifier in the in-processing stage, as in-processing methods have demonstrated good flexibility in both balancing fairness and supporting multiple classifiers and fairness metrics.

2.3 Interaction between Privacy and Fairness

Privacy and fairness interact closely with each other not only because some tasks involve privacy and fairness issues together, but also because implementing one usually will have an impact on the other. In the following, we will introduce existing work in these two lines.

2.3.1 Simultaneous Enforcement of Privacy and Fairness

Several papers have studied the privacy and fairness issues in machine learning simultaneously. Differentially private and fair logistic regression models are proposed in [40, 139]. Jagielski et al. studied fair learning under the constraint of differential privacy with equal odds [67]. In [38], a notion of approximate fairness is proposed in the finite sample access setting and presents a private PAC learner that is differentially private and satisfies approximate fairness with high probability. In [115], a new variational approach is proposed to learn private and fair representations which are based on the Lagrangians of a new formulation of the privacy and fairness optimization problems.

2.3.2 Mutual Impact of Privacy and Fairness

The impact of DP-SGD on model accuracy has been considered in [9, 138]. Pannekoek et al. explored the impact of differential privacy and fairness constraints on the privacy-utility-fairness in different conditions [107]. The impact of differential privacy is studied on fair and equitable decision-making [113]. Chang et al. studied privacy risks of group fairness through the lens of membership inference attacks, and find that fairness comes at the cost of privacy, and this cost is not distributed equally [22]. Recent studies [21, 101, 103] have demonstrated that model fairness can be skewed by data poison attacks. Mehrabi et al. [101] and Chang et al. [21] studied data poisoning attacks against group-based fair machine learning. Nanda et al. proved that it might be easier for an attacker to target a particular group, resulting in a form of robustness bias [103].

CHAPTER 3

Correlated Differential Privacy: Feature Selection in Machine Learning

In the first main chapter of the thesis, we work on a privacy issue in machine learning. Machine learning requires a large amount of data, and data are increasing collected from human activities. In some scenarios, data examples have correlations and correlated data used for industrial applications can disclose more private information in machine learning algorithms when using differential privacy. This chapter presents a correlation reduction scheme based on feature selection that helps to improve data utility when applying differentially private machine learning algorithms.

3.1 Introduction

Machine learning becomes an indispensable and efficient tool to provide services for human beings in industrial applications, such as the Internet of Things (IoT) [119] and smart cities [61]. One main data source used for machine learning in the industry is from human activities. For example, human data are often collected via smartphones and these data are analyzed to provide some services in smart cities, such as traffic monitoring [143] and smart health [123]. Data collected from humans usually contain some sensitive information, such as location information and health data in the above examples. When these data are used for machine learning, individual privacy can be leaked [55].

As a popular technique for privacy-preserving, differential privacy was first proposed by Dwork et al. [47]. Since then, differential privacy has attracted considerable attention because it provides a rigorous mathematical framework for preserving privacy. Recently, differential privacy is widely used to protect privacy in industrial informatics, such as location privacy protection [141] and smart grids [51], [152]. Many works have addressed the privacy issue in machine learning with differential privacy. For example, Chaudhuri et al. provided an output perturbation where the model is trained and then adds noise to the output [23] and an objective perturbation mechanism where a carefully designed linear perturbation item is added to the original loss function [25]. Yang et al. proposed a differentially private feature selection scheme [140] and Abadi et al. considered the privacy issue with differential privacy in deep learning [1]. However, previous works have not considered the data correlation when designing differentially private machine learning algorithms.

In the definition of differential privacy, data within a dataset are assumed to be independent. This is a somewhat faulty assumption since data in industrial applications are always correlated beginning from when the data is first generated, such as temporal datasets in monitoring systems. Intuitively, when some of the records in a dataset are correlated, deleting one record may have a great impact on the other records, which could reveal more information to an adversary than expected. Kifer and Machanavajjhala's study on data correlations [76] confirms this observation, and the finding has launched a new stream of research on how to preserve privacy in correlated datasets when they are released to the public. For example, Liu et al. introduced dependent differential privacy with a defined parameter, called dependence coefficient, to describe data correlation [88]. A second privacy framework was designed, called Pufferfish, which is flexible and can provide a privacy guarantee for various data sharing needs [77].

Overall, the contributions of this chapter can be summarized as follows:

- 1) We proposed a differentially private feature selection based on feature importance. The proposed method can select features privately while retaining a desirable data utility.

- 2) We propose a correlation reduction scheme based on feature selection to reduce data correlation in correlated datasets. This helps to reduce the correlated sensitivity when implementing differentially private machine learning algorithms, and thus improves data utility.

- 3) Experiments validate the effectiveness of our proposed feature selection scheme. The results show improved data utility for both data analysis and data publishing.

3.2 Preliminaries

3.2.1 Differential Privacy

Differential privacy is a rigorous privacy model [47]. In brief, given two datasets D and D' that contain a set of records, these are referred to as neighbouring datasets when they differ in one record. Let D denote a dataset with a number of N data records (or data examples), and each data record has a number of K features. A query q is a function that maps the record $r \in D$ into outputs $q(D) \in \Omega$, where Ω is the whole set of outputs.

Definition 3.1. (ϵ-Differential privacy) [47] Given neighboring datasets D and D' that differ in one data point, an algorithm $\mathscr{M}$ satisfies (ϵ)-differential privacy for any possible outcome Ω

$$Pr[\mathscr{M}(D) \in \Omega] \leq \exp(\epsilon) \cdot Pr[\mathscr{M}(D') \in \Omega], \tag{3.1}$$

where ϵ is privacy budget which determines privacy level.

Definition 3.2. (Sensitivity) [47] For a query $q : D \to R$, and neighboring datasets D and D', the sensitivity of q is defined as

$$\Delta q = \max_{D,D'} ||q(D) - q(D')||. \tag{3.2}$$

where $||\cdot||$ denotes L_1 or L_2 norm. Sensitivity measures the maximal difference between neighboring datasets.

Definition 3.3. (Laplace mechanism) [48] For any query q: $D \to R$ over a dataset D, the following mechanism provides ϵ-differential privacy if

$$\mathscr{M}(D) = q(D) + Laplace(\Delta q/\epsilon) \tag{3.3}$$

The Laplace noise is denoted as $Laplace(\cdot)$ and is drawn from a Laplace distribution with the probability density function $p(x|\lambda) = \frac{1}{2\lambda}e^{-|x|/\lambda}$, where λ relate to the sensitivity and the privacy budget.

Theorem 3.1. ***(Sequential composition)*** *[48] Suppose that a set of privacy mechanisms $\mathscr{M}$={$\mathscr{M}_1, ..., \mathscr{M}_m$}, gives ϵ_i differential privacy ($i = 1, 2..., m$), and these mechanisms are sequentially performed on a dataset. $\mathscr{M}$ will provides ($\sum_i \epsilon_i$)-differential privacy for this dataset.*

3.2.2 Feature Selection

Feature selection is a method for selecting the attributes in a dataset (such as columns in tabular data) that are most relevant to the prediction [20]. In other words, feature selection largely acts as a filter that sifts out features that are less useful to solving a problem. With feature selection, both the efficiency and the accuracy of the predicted results can be improved.

In this chapter, we adopt feature importance to select features. Feature importance is a method of ranking features based on random forests. Feature importance is measured according to the mean decrease in impurity, which is defined as the total decrease in node impurity averaged over the forest. This score can be computed automatically for each feature after training and scaling the results so that the sum of importance for all features is equal to 1. One strength of the random forest is that it is easy to measure which features are relatively more important to the results. With this method, we are able to select the most important features in the dataset.

3.3 Example of Traffic Monitoring

In this section, we present the issue of correlated data in differential privacy with a detailed industrial example of traffic monitoring and show how correlated data can degrade the level of privacy in industry applications.

The traffic monitoring is one of most used technologies in smart cities. Location information of users in a region are collected by a trusted server and the aggregate information of the dataset (i.e., the counts of users at each location) is continuously released to the public. Some users in the region may have a form of social relationship, perhaps family members. In this case, some users may have the same location information during some time and hence the records of user information can be correlated in the dataset.

As shown in Table 3.1, the user's locations are recorded at different time points. It is assumed that users only appear in one location at each time point, and it is observed that $user_1$ and $user_2$ take the same route from time point $t = 1$ to $t = 4$ (they may have social relationships). In this case, if one were to change the location of $user_1$, the location of $user_2$ would also change. In this way, the records for $user_1$ and the records for $user_2$ are correlated.

Table 3.1: Users' locations at different times

t user	1	2	3	4
u_1	loc_2	loc_2	loc_3	loc_4
u_2	loc_2	loc_2	loc_3	loc_4
u_3	loc_1	loc_4	loc_5	loc_2
u_4	loc_4	loc_5	loc_2	loc_5

Table 3.2: The sum counts of users' locations

t loc	1	2	3	4
loc_1	1	0	0	0
loc_2	2	2	1	1
loc_3	0	0	2	0
loc_4	1	1	0	2
loc_5	0	1	1	1

Table 3.2 shows that the some counts at different locations are always 2. In terms of the Laplace mechanism, adding the amount of $Lap(1/\epsilon)$ noise to perturb each count in Table 2 can achieve ϵ-DP at each time point. However, the expected privacy guarantee may breach with correlated records in the dataset. With background information of who has the relationship in a certain region, an attack can infer the location information of $user_1$ and $user_2$ at different time points. Consequently, after releasing private count of user's locations, the location information of $user_1$ and $user_2$ may not be ϵ- differentially private as expected. Instead, it is 2ϵ-differentially private since changing one user's location will change the count 2.

In summary, this example shows that correlated data in a dataset will disclose more information than expected when these data are used for machine learning algorithms in industrial applications. Essentially, adding more noise to a correlated dataset is a way to guarantee differential privacy. Such a case reveals the level of challenge in industries when dealing with correlated data in situations where differential privacy must be satisfied, but high-quality query results must be maintained.

3.4 The Extent of Data Correlation

3.4.1 Correlated Degree

Inspired by [155], we have incorporated the notion of correlated degree $\theta_{ij} \in [-1,1]$ to denote the extent of correlation between record i and record j. When $|\theta_{ij}| > 0$, record i and record j have a positive correlation and vice versa. When $|\theta_{ij}| = 1$, record i and record j are fully correlated and When $\theta_{ij} = 0$, there is no relationship. When there are a number of l records in a dataset, it is possible to list the relationship for all records and form a correlated degree matrix Λ.

$$\Lambda = \begin{pmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1N} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{N1} & \theta_{N2} & \cdots & \theta_{NN} \end{pmatrix} \tag{3.4}$$

A threshold θ_0 is defined so as to select strongly correlated records. For a given θ_0, the value of the correlated degree is

$$\theta_{ij} = \begin{cases} \theta_{ij}, & \theta_{ij} \geq \theta_0, \\ 0, & \theta_{ij} < \theta_0, \end{cases} \tag{3.5}$$

A correlated degree matrix can describe the correlations of the whole dataset and, once analyzed, the curator will hold all knowledge of the data correlations. Data privacy can still be protected, even when the adversary is privy to the entire correlated degree matrix, if enough noise is added to mask the highest impact of deleting one record using correlated differential privacy.

3.4.2 Correlated Sensitivity

Global sensitivity can only measure the maximal number of correlated records but does not consider the extent of the data correlation. Hence, the notion of correlated sensitivity is introduced to measure the extent of the impact on other records from changing one record. As mentioned earlier, global sensitivity adds extra noise by simply multiplying the maximal number of correlated records. Whereas, correlated sensitivity is able to model the correlations in a more exact way.

Definition 3.4. (Correlated sensitivity) For a query q, correlated sensitivity is based on the correlated degree and the number of correlated records, which is defined as

$$\Delta CS_q = \max_{i \in \amalg} \sum_{j=0}^{N} |\theta_{ij}| \{ \|(q(D^j) - q(D^{-j})\|_1 \} \tag{3.6}$$

where ⨿ is the set of records in a dataset, and θ_{ij} is the correlated degree between record i and record j. D_j and D_{-j} are neighboring datasets that differ by record j. Correlated sensitivity lists all the sensitivity of records with the query q. With correlated sensitivity, the maximal effect on all records of a dataset can be measured when one record is deleted. For any query q, the perturbed answer is calibrated with the equation,

$$\hat{q}(D) = q(D) + Laplace(\frac{\Delta CS_q}{\epsilon}) \tag{3.7}$$

For any query q, the correlated sensitivity is smaller than the global sensitivity. The global sensitivity is denoted as $\Delta GS_q = \max_{i \in \amalg} \sum_{j=0}^{l} \{ l \, \|(q(D^j) - q(D^{-j})\|_1 \}$, where l denotes the number of correlated records. Since we use the correlated degree $\theta_{ij} \in [-1,1]$ to describe the extent of data correlation, the correlated sensitivity is no larger than the global sensitivity.

We note that the correlated degree θ_{ij} is related to every feature in record i and record j. When deleting features in the dataset, the extent of correlation between record i and record j will also be changed. Thus, after describing the extent of data correlation in a dataset, we use feature selection to reduce data correlation.

3.5 Correlation Reduction Based on Feature Selection

3.5.1 Overview of the Method

In our method, we select features in terms of three principles: 1) the accuracy of training results; 2) the privacy of feature selection; 3) the reduction of the data correlation. As Fig. 1 shows, the proposed correlation reduction based on feature selection scheme (CR-FS) involves five steps: 1) removing collinear features; 2) removing unimportant features; 3) choosing features with differential privacy; 4) obtaining the useful feature set $\mathscr{B}$; and 5) adjusting the features that can reduce data correlation within the dataset. Each of these methods is described in detail in the following sections.

3.5.2 The Proposed CR-FS Scheme

Following traditional feature selection, we propose the Algorithm I that selects features with differential privacy. For a given dataset, feature selection is a crucial step before executing a machine learning algorithm, especially with high-dimensional datasets. Many features in the dataset are entirely irrelevant, insignificant, or are collinear, which will have a negative impact on the result. Additionally, retaining more features typically leads to a higher degree of data correlation, which, with differential privacy, negatively impacts the privacy level. Hence, our goal is to select a subset of features with relatively lower levels of data correlation while maintaining good utility for data publishing and analysis.

3.5.2.1 Removing Collinear Features

The first step is to filter out the collinear features that can decrease generalization performance on the test set due to less model interpretability and high variance. Usually, the extent of collinearity between features is calculated by the absolute magnitude of the Pearson's correlation coefficient. The calculation of Pearson's correlation coefficient is

$$\rho_{m,n} = \frac{E[(f_m - \mu_{f_m})(f_n - \mu_{f_n})]}{\sigma_{f_m}\sigma_{f_n}} \tag{3.8}$$

Algorithm 1 Differentially private feature selection scheme

```
Input: Dataset, $T_{cf}$, $T_{fi}$,$T_{mv}$, $\epsilon_1$;
Calculate feature collinearity according to Eq. 3.5;
if $\rho_{m,n} \leq T_{cf}$ then
    Remove $f_m$ or $f_n$;
end if
Remove unimportant features with $T_{fi}$;
Remove missing values with $T_{mv}$
Calculate the $h_k$ by Random forest;
Calculate the sensitivity $\Delta h$ according to Eq. (3.10);
for $h_k \in \{h_1,...,h_K\}$: do
    Add Laplace noise $\hat{h}_k = h_k + Lap(\frac{\Delta h_q}{\epsilon_1})$;
end for
Do the normalization $h_k = \hat{h}_k / \sum_{k=1}^{K} \hat{h_k}$;
for $h_k \in \{h_1,...,h_K\}$: do
    Delete features one by one according to the sequence of feature importance;
    Calculate model predictions;
end for
Find the useful feature set $\mathscr{B} = \{f_1, f_2, ...f_k\}$ and the adjusted feature set: $\mathscr{A} = \{f_{k+1},...,f_K\}$;
Add or delete features from the adjust feature set $\mathscr{A}$ according to Algorithm 2;
Output: Useful feature set $\mathscr{B}$, an adjusted feature set $\mathscr{A}$;
```

Where f_m and f_n are two random features in the dataset; μ_{f_m} and μ_{f_n} are the mean of f_m and f_n; σ_{f_m} and σ_{f_m} are the standard deviation of features f_m and f_n. In our scheme, we set a threshold of $T_{cf} \in [0,1]$ to identify collinear features and remove the features with a collinearity of greater than T_{cf}.

3.5.2.2 Removing Unimportant Features

The second step is to remove unimportant features, including 1) features of zero importance and features of low importance; 2) features with a high percentage of missing values; and 3) features with a single value. Zero and low importance features can be identified using the feature importance threshold, denoted as $T_{fi} \in [0,1]$. Features with an importance value of lower than T_{fi} will be removed. The threshold for missing values is defined as $T_{mv} \in [0,1]$, and features with a percentage of missing values greater than T_{mv} will be removed.

3.5.2.3 Choosing Features with Differential Privacy

We adopt feature importance h in Random forest to calculate the feature weight for each feature. Neighboring data is obtained when record r_i is deleted, the feature importance can be calculated by Random forest and the feature importance $h_1^i, h_2^i, ..., h_N^i$ are sorted in an increasing order. Based on this, we introduced the notion of record sensitivity of feature importance.

Definition 3.5. (Record sensitivity of feature importance) For a query q, the record sensitivity of feature importance of r_i can be defined as,

$$\Delta h_k = ||h_k^i - h_1^i||_1 \tag{3.9}$$

Definition 3.6. (Sensitivity of feature importance) For a query q, the sensitivity of feature importance is determined by the maximal record sensitivity of feature importance,

$$\Delta h_q = \max_{i \in \amalg}(\Delta h_k) \leq 1 \tag{3.10}$$

where $\amalg$ is a set of records related to a query q. It is easy to know the sensitivity of feature importance is $\Delta h_q \leq 1$, since the range of feature importance is from 0 to 1. We apply Laplace mechanism to add noise to the feature importance. The perturbed feature importance can be denoted as,

$$\hat{h}_k = h_k + Lap(\frac{\Delta h_q}{\epsilon}) \tag{3.11}$$

Since the sum of the feature importance $\sum_{k=1}^{K} h_k = 1$, we normalize the perturbed feature importance as follow,

$$h_k = \hat{h}_k / \sum_{k=1}^{N} \hat{h}_k \tag{3.12}$$

The new sequence of feature importance can be denoted as $h_1 < h_2 < ... < h_K$.

3.5.2.4 Select Features Based on Model Accuracy

The third step is to find the useful feature set. The useful feature set $\mathscr{B}$ contains the features that will produce the best prediction results by the machine learning algorithm. In our method, the less important features are deleted one by one in the order of feature importance until the best chance of accurate predictions is achieved. Practically, finding useful feature set with this method demands far less computational overhead than other methods. The features that have not been selected for the useful feature set are stored as the adjusted feature set. These features can be used later for a tradeoff between utility and privacy. The useful feature set $\mathscr{B}$ can be denoted as $\mathscr{B} = \{f_1, f_2, ..., f_k\}$ and the adjusted feature set $\mathscr{A}$ can be denoted as $\{f_{k+1}, f_{k+2}, ..., f_K\}$.

3.5.2.5 Adjust Features Based on Data Correlations

The final step is to adjust some features based on the useful feature set $\mathscr{B}$ in order to reduce data correlation over the whole dataset, as a way to balance the tradeoff between utility and correlated sensitivity. Basically, the correlated sensitivity of a dataset is irrelevant to the number of features. This means that more features of a dataset may have a lower correlated sensitivity and less features may have a higher correlated sensitivity. Useful feature set $\mathscr{B}$ can achieve a good data utility without privacy guarantee, yet it may have a higher correlated sensitivity and a high correlated sensitivity has a huge impact on utility for data publishing and data analysis. In other words, if the goal is to generate a differentially private dataset with good data utility, the process of feature selection should also consider correlated sensitivity.

Algorithm 2 shows the adjusted feature selection scheme, which includes backward and forward feature selection methods. The forward feature selection adds features one by one from the adjusted feature set $\mathscr{A}$ to useful feature set $\mathscr{B}$. The correlated sensitivity is calculated according to Eq. 3.6, and then Laplace noise is added according to Eq. 3.7. Training with these added features can obtain the feature set $\mathscr{A}_1$, which provides optimal performance. However, sometimes adding a large number of features only slightly increases performance, particularly with high dimension datasets, while too many features can lead to a less interpretive model. Hence, when a set of added features appears to be more or less equally good, then it makes sense to choose the simplest feature set. We set a threshold T to evaluate the difference of training results.

Algorithm 2 Adjusted feature selection scheme

```
Input: Useful feature set $\mathscr{B}$, adjusted feature set $\mathscr{A}$, $\epsilon_2, \theta_0$;
for $f_k \in \{f_{k+1}, ..., f_K\}$: do
    Add features to the useful feature set $\mathscr{B}$ from the adjusted feature set $\mathscr{A}$;
    Calculate the correlated sensitivity of new datasets according to Eq. 3.6;
    Add Laplace noise $Lap = \frac{\Delta CS_q}{\epsilon_2}$;
    Train on the dataset and get predicted results;
end for
Obtain the adjusted feature set $\mathscr{A}_1$ that has the best performance;
for $f_k \in \{f_1, ..., f_k\}$: do
    Delete features from the useful feature set $\mathscr{B}$ one by one;
    Calculate the correlated sensitivity of new datasets according to Eq. 3.6;
    Add Laplace noise $Lap = \frac{\Delta CS_q}{\epsilon_2}$;
    Train on the dataset and get predicted results;
end for
Obtain the adjusted feature set $\mathscr{A}_2$ that has the best prediction;
if $s(\mathscr{A}_1) \geq s(\mathscr{A}_2)$ then
    $\mathscr{A}_1$ is the adjusted feature set $\mathscr{A}$;
else
    $\mathscr{A}_2$ is the adjusted feature set $\mathscr{A}$;
end if
Output: Adjusted feature set $\mathscr{A}$;
```

If the difference of training results is smaller than T, we select the simplest feature set that has the smallest number of predictors.

In backward feature selection, features in set are deleted one by one according to their feature importance. By comparing the training results with different deleted features, feature set $\mathscr{A}_2$ is generated, which has the best performance. Similar to forward feature selection, when a set of deleted features appears to be more or less equally good, it makes sense to choose the simplest feature set. We also use the threshold T to select the simplest feature set. Ultimately, the adjusted feature set $\mathscr{A}$ is determined by comparing the training result $s(\mathscr{A}_1)$ and $s(\mathscr{A}_2)$.

3.5.3 Discussion

useful feature subset $\mathscr{B}$ and adjusted feature set $\mathscr{A}$, represent the balance between utility and correlated sensitivity. Adding the adjusted features is likely to degrade data utility somewhat, but these extra features serve to reduce the correlated sensitivity of the dataset, which offsets the reduction in utility. The overall result is a feature selection scheme that strikes a balance that leads to less data correlation while maintaining good data utility for data analysis and data publishing.

Our proposed scheme has three advantages. First, feature importance is a computationally efficient method for generating the useful feature set compared to some of the other existing methods. Feature importance is the variable that provides the guide to select which features are best to add or delete. Second, with differential privacy, we can choose features privately. Third, with the consideration of data correlation, we can select features that has less data correlation in the whole dataset and thus reduce the correlated sensitivity and improve the data utility of the dataset.

3.6 Experiments

Our evaluation experiments involve four real-world datasets in terms of both data analysis and data publishing tasks [153]. Utility for data analysis is tested with two machine learning algorithms: LR and linear SVM. Utility for data publishing is tested on count and mean queries.

3.6.1 Experimental Setup

3.6.1.1 Dataset

The experiments involve four datasets, which have different extent of data correlation and different number of features.

- Adult Dataset: Adult Dataset is from the UCI Machine Learning repository. After data preprocessing, we extract 3000 records with 12 features.
- Breast cancer Dataset: This dataset can be found on UCI Machine Learning Repository. After data preprocessing resulted in 569 records with 20 features.
- Titanic Dataset: This dataset comes from a Kaggle competition where the goal was to analyze which sorts of people were likely to survive the sinking of the Titanic. After data preprocessing, we extract 891 records with 9 features.

- Porto Seguro Dataset: Porto Seguro is a well-known auto and homeowner insurance company. After preprocessing, we extract 1770 records with 37 features.

3.6.1.2 Comparison

For better comparisons, four schemes are considered in the experiments.

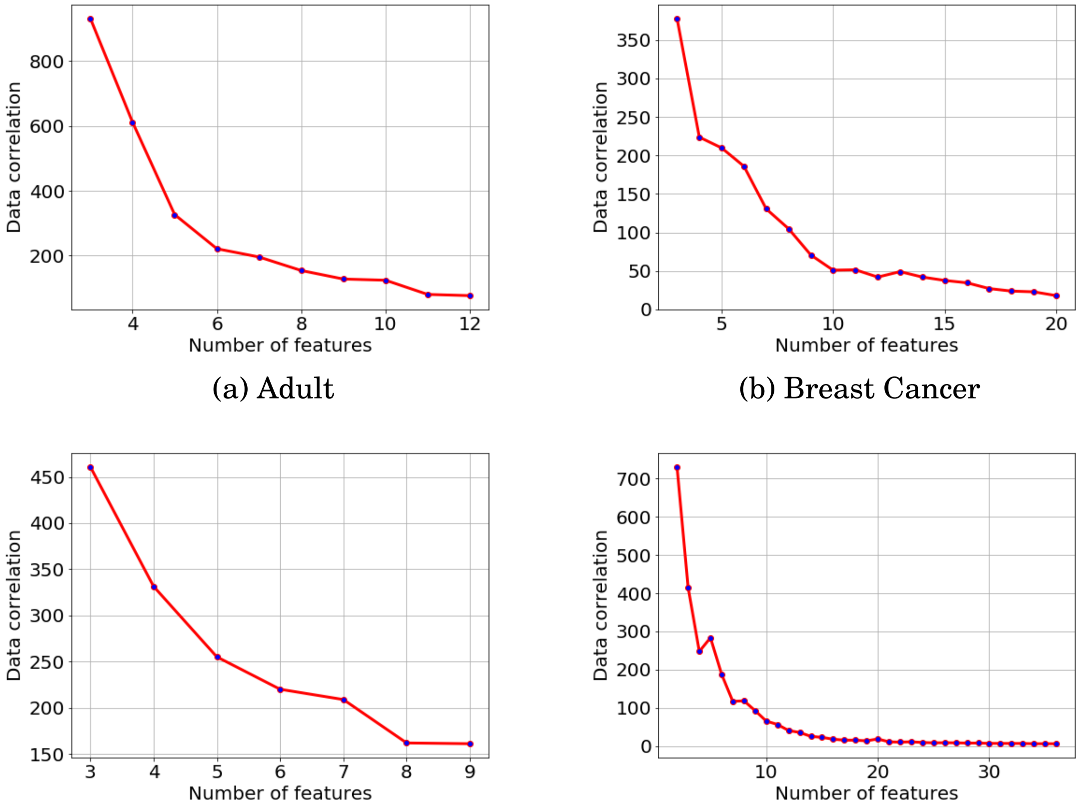
- A non-private scheme, where the dataset has no privacy protection.
- The group scheme, where noise is added by multiplying the number of correlated records, as proposed by Chen et al. in [27].
- The Zhu scheme, where noise is added according to the correlated sensitivity [155].
- The proposed scheme, where noise is added according to the CR-FS scheme defined in this chapter.

3.6.1.3 Parameters

For correlation knowledge between records, no dataset suggests pre-defined knowledge of any correlated data. We use Pearson correlation coefficient to construct the correlated degree matrix, where a correlation exists for record i and record j if $\theta_{ij} \geq \theta_0$. θ_0 is set to 0.9 for Adult Dataset, Breast cancer Dataset and Breast cancer Dataset and θ_0 in Porto Seguro Dataset is set to 0.7. For correlation knowledge between features, the Pearson correlation coefficient threshold T_{fi} is set to 0.9. The missing value threshold T_{mv} is set to 0.2. The threshold of feature importance T_{fi} is set to 0.9.

Table 3.3: Number of features in different stages

	Original dataset	After data preparation	$\mathscr{B}$	$\mathscr{A}$
Adult	15	12	8	12
Breast cancer	32	20	10	17
Titanic	12	9	7	9
Porto seguro	59	37	14	28



(a) Adult

(b) Breast Cancer

(c) Titanic

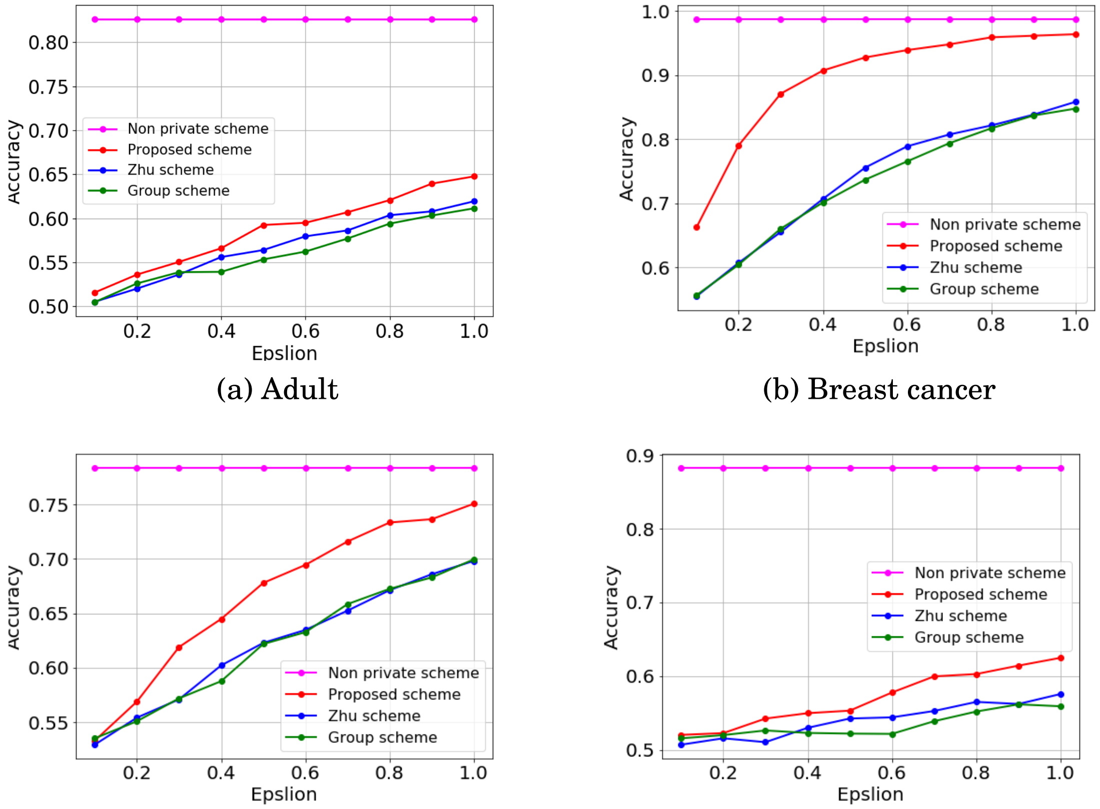
(d) Porto Seguro

Figure 3.1: Data correlation for different number of features

3.6.2 Experiments for Data Analysis

One aim of our proposed scheme is to improve utility for data analysis, which we evaluate according to the accuracy of the predicted results. For this set of experiments, we choose two machine learning algorithms - LR and linear SVM - and test the output perturbation to assess data utility.

Fig. 3.1 shows that, according to the Pearson correlation coefficients, data correlation varies with the number of features. Data correlation generally decreases with a growing number of features but eventually stabilizes. For example, Figs. 3.1b and 3.2c show that data correlation becomes stable at 17 features with the Breast Cancer dataset and at 8 features with the Titanic dataset. This observation indicates that data correlation across the entire dataset can be reduced while preserving a suitable number of features



(a) Adult

(b) Breast cancer

(c) Titanic

(d) Porto Seguro

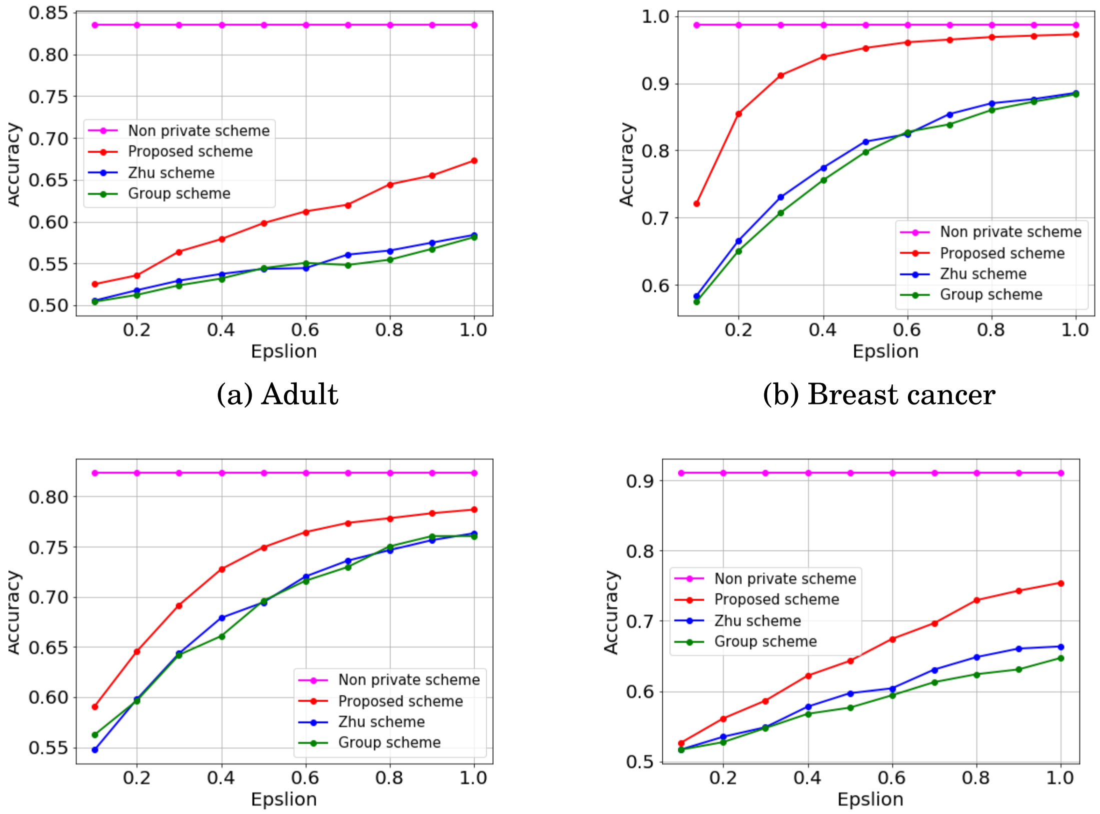
Figure 3.2: Privacy-Accuracy trade-off in SVM for different datasets

for data analysis because more features mean less correlation.

Table 3.3 shows the number of features in each dataset at different stages of the proposed scheme. It is noted that useful feature set $\mathscr{B}$ will always contain more features than the adjusted feature set $\mathscr{A}$ and, as shown in the table, the adjusted feature set $\mathscr{A}$ have less data correlation than the useful feature set $\mathscr{B}$, demonstrating that more features reduce correlation in a correlated dataset.

Figs. 3.2 and 3.3 show the performance of linear SVM and LR on different datasets with the four schemes. In most cases, LR has better accuracy than linear SVM. For example, Fig. 3.2 (a) shows that when $\epsilon = 1$, LR have an accuracy of around 0.675 versus linear SVM's 0.645. Accuracy with the non-private scheme remains constant as the privacy budget increases and also performed better than the other schemes. This result demonstrates that imposing any form of privacy requirement on a dataset degrades data utility.

For the private schemes, the proposed scheme outperforms both the group and Zhu schemes in all circumstances. Figs. 3.2 and 3.3 show the level of improvement, especially Fig. 3.2 (b). $\epsilon = 1$, the proposed scheme scores an accuracy of around 0.97 compared to around 0.85 for the Zhu scheme. We attribute the improved performance of our scheme to the adjusted features. These additional features reduce data correlation but have little impact on the prediction results. Less data correlation means less noise needs to be added, which leads to better data utility. Other schemes do not reduce data correlation; they only consider how to accurately describe the data correlations, without considering that data correlation actually impedes accuracy.



(a) Adult

(b) Breast cancer

(c) Titanic

(d) Porto Seguro

Figure 3.3: Privacy-Accuracy trade-off in LR for different datasets

Additionally, baselines present closed curves with the first three datasets because the Pearson coefficient is set to a high-value $\theta_0 = 0.9$. This results in a similar correlated sensitivity for both schemes and, consequently, a similar level of noise is added. However, with the Porto Seguro dataset, we set the Pearson coefficient to $\theta_0 = 0.7$. Hence, there is a minor gap in performance. Also worthy of note is that the accuracy of prediction results varied for different datasets. This is due to the amount of data correlation in each dataset; higher correlation means more noise must be added, which reduces accuracy.

3.6.3 Experiments for Data Publishing

The second aim of our scheme is to improve the utility for data publishing, which we evaluate with both count and mean queries. Mean absolute error (MAE) is used as the metric to assess both results, but different calculation formulas are defined to analyze the base results and the impact of varying the privacy budget. The accuracy of common queries is measured by MAE, which is given as,

$$MAE = \frac{1}{|q|} \sum_{q_i \in q} |\hat{q}_i(x) - q_i(x)| \tag{3.13}$$

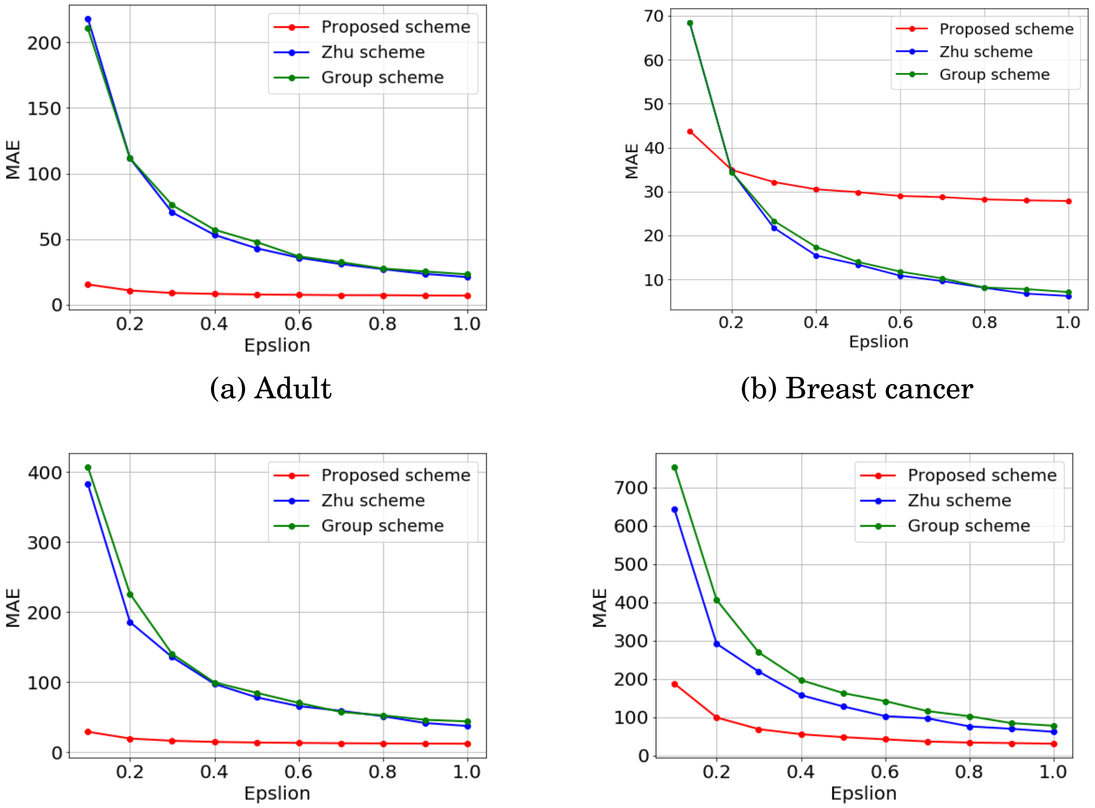
where $q_i(x)$ is the true aggregation result for one query, and $\hat{q}_i(x)$ is the perturbed aggregation result calculates through different schemes. A low MAE indicates a low error and, thus, a better data utility.

To analyze how the proposed scheme performs with different privacy budgets, we also define a second MAE containing two components. One component measures the noise added due to correlated sensitivity, and the other measures the errors introduced by adding the adjusted features. These features have an impact on a new query object that can emerge as errors when comparing the adjusted dataset to the original. This MAE is defined as

$$MAE = \frac{1}{|q|} \sum_{q_i \in q} |\hat{q}_i(x) - (q_i(x) - q_i^o(x))| \tag{3.14}$$

where $\hat{q}_i(x)$ and $q_i^o(x)$ are the true aggregation result on the useful feature set $\mathscr{B}$ and the adjust feature set $\mathscr{A}$, respectively.

Fig. 3.4 shows the impact of varying privacy budgets on the performance of count queries in terms of MAE. With the proposed scheme, the MAE decreases as the privacy



(a) Adult

(b) Breast cancer

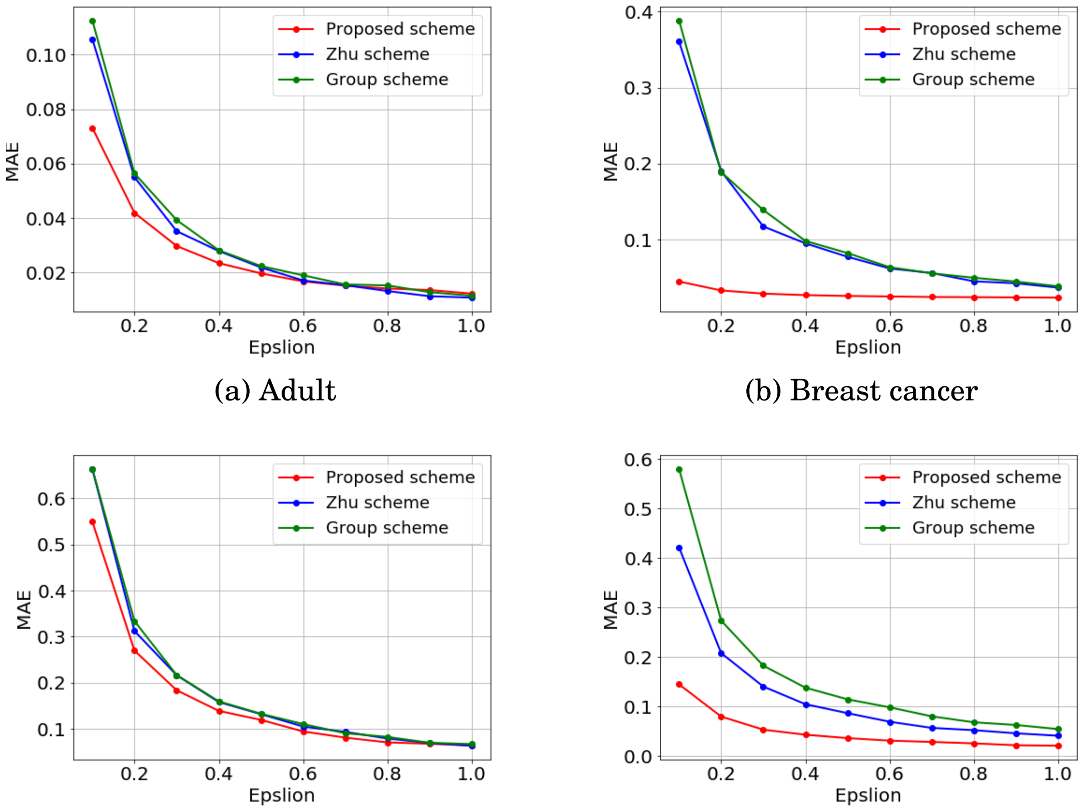
(c) Titanic

(d) Porto Seguro

Figure 3.4: MAE performance for count queries

budget grows before stabilizing toward the end. This result demonstrates that a lower privacy requirement has better data utility. Moreover, the MAE for the proposed scheme is significantly smaller than the other schemes, which means that the proposed scheme does indeed improve data utility. For example, Figs. 3.4 (a) and 3.4 (b) at $\epsilon = 0.2$ show an MAE of around 18 for the Adult dataset and 17 for the Titanic dataset using our proposed scheme, whereas the group and Zhu schemes return an MAE of around 110 and 200 on these same datasets - an enormous increase over the proposed scheme. Again, we attribute these results to reduced data correlation after adding the adjusted features.

In terms of the other schemes, the MAE for the Zhu scheme is slightly lower than for the group scheme most of the time for the same reason as explained in the data



(a) Adult

(b) Breast cancer

(c) Titanic

(d) Porto Seguro

Figure 3.5: MAE performance for mean queries

analysis experiments. Moreover, the MAE for the Zhu scheme decreases faster as the privacy budget increases from 0.1 to 0.4 than when the budget increases from 0.4 to 1. This again shows that a higher privacy requirement creates a higher data utility cost.

The results of varying the privacy budgets with mean queries are similar, as shown in Fig. 3.5. However, the MAE is much smaller than for the count queries. This is because, after data normalization, the scale of data falls within $[-1,1]$; therefore, each record has a similar mean value. As a result, the outcomes of mean queries are much smaller than for count queries. In addition, the MAE for our proposed scheme is not always better than the group or Zhu schemes - for example, when $\epsilon < 0.2$. This shows that adding the adjusted features can introduce additional errors. Hence, the quality of the query results in the proposed scheme depends on the type of queries and the dataset itself but, overall, our proposed scheme returns a lower MAE than the other schemes.

3.6.4 Discussion

The key to the CR-FS scheme is to reduce data correlation in the whole dataset while maintaining a good utility for data analysis and data queries. We add differential private noise in two places: feature selection and data training and still can achieve desirable performance. This is because the fact that sensitivity of feature selection is smaller than 1, and the sequence of feature importance will not change much. That is to say, there is a high probability that more important features are still more important and less important features are still less important. In this way, a higher probability that important features are kept for training and less important features are used to reduce data correlation.

For data analysis, we select features in step 5 according to the accuracy of predicted results, thus the selected features can have less correlation across the whole dataset and achieve desirable accuracy results. For data queries, the correlation in the whole dataset is also reduced with the proposed CR-FS scheme. However, as we noted in Figures 3.4 and 3.5, the MAE is not always better than other schemes. This is because the sensitivity is related to the type of queries and the dataset itself. Deleted or added features in the dataset can reduce the data correlation, which may bring in a new error with regard to different queries.

3.7 Summary

In this chapter, we identified the privacy issue of the data correlation in machine learning, which may result in more privacy loss than expected in industrial applications. We propose a novel feature selection scheme CR-FS to reduce data correlation with little compromise to data utility. The proposed CR-FS scheme includes steps that consider the accuracy of predicted results, privacy-preserving and the data correlation in the dataset. Our proposed algorithm strikes a better trade-off between data utility and privacy leaks for correlated datasets. The method's performance is evaluated via extensive experiments, and the results prove that our proposed CR-FS scheme provides better data utility for both data analysis and data queries compared to previous schemes.

CHAPTER 4

FAIRNESS IN SEMI-SUPERVISED LEARNING: UNLABELED DATA HELP TO REDUCE DISCRIMINATION

Machine learning not only has privacy problems, but also may produce discrimination. Fairness in machine learning has become a very important issue as we move forward in the world of machine-assisted prediction for humans. The ability to design machine learning algorithms that treat all groups equally may be one of the most influential factors in allocating resources and opportunities to humans. In this chapter, we focus on fairness in semi-supervised learning and investigate if unlabeled data help to reduce discrimination. More specifically, we propose a fairness-enhanced sampling framework that combines pseudo labeling, re-sampling and ensemble learning to achieve fair semi-supervised learning in the pre-processing phase.

4.1 Introduction

Machine learning is now in wide use as a decision-making tool in many areas, such as job employment, risk assessment, loan approvals and many other basic precursors to equity. However, the popularity of machine learning has raised concerns about whether the decisions algorithms make are fair to all individuals. For example, Chouldechova found evidence of racial bias in recidivism prediction tool where black defendants are more likely to be assessed with high risk than white defendants [31]. Obermeyer et al. found prejudice in health care systems where black patients assigned the same level of risk by the algorithm are sicker than white patients [105]. These findings show that unfair machine learning algorithms will affect legal justices, health care, and other aspects of human beings.

Over the past few years, much research has been devoted to designing fairness metrics, such as statistical fairness [15, 31, 146], individual fairness [45, 72, 92] and causal fairness [78, 83]. These approaches and algorithms can be roughly divided into three categories: pre-processing methods, in-processing methods and post-processing methods. Pre-processing methods adjust data distribution [15, 74] or learn new fair representations [94, 124, 147], to relieve some of the tension between accuracy and fairness. In-processing methods add constraints or regularizers to restrict the correlation between labels and sensitive/protected attributes, i.e., traits that can be targeted for discrimination [75, 81, 146]. Post-process methods calibrate training results [60]. These studies mainly focus on addressing the two most crucial fundamental issues in machine learning fairness: how to formalize the concept of fairness in the context of machine learning tasks, and how to design effective algorithms to achieve an ideal compromise between accuracy and fairness.

However, almost all methods of achieving fairness are mostly for either supervised learning or unsupervised learning, and fair semi-supervised learning (SSL) has rarely been considered. Realistically though, training data is often a combination of labeled and unlabeled samples, so a semi-supervised solution has high practical value. Also, since the ideal is a lofty goal, the trade-off between accuracy and fairness is still an ongoing pursuit. [26] showed that increasing the amount of training data is likely to produce a better trade-off between accuracy and fairness. This insight inspired us to wonder whether using unlabeled data to augment the training set might give us a kind of control value with which to balance fairness and accuracy. Unlabeled data is abundant and, if it could be used as training data, we could adjust the size of the training set as required to meet accuracy vs. fairness thresholds. We may even be able to avoid the need to make a compromise between fairness and accuracy entirely. This leaves fair semi-supervised learning with two challenges: 1) How to make use of unlabeled data to achieve a better trade-off between accuracy and fairness; and 2) How to alleviate the impact of noise, which is common to semi-supervised learning.

To tackle these challenges, we propose a framework to achieve fair SSL in the preprocessing phase. The solution to the trade-off challenge is to use unlabeled data to reduce representation discrimination. (Representation discrimination is due to certain parts of the input space under-represented.) Therefore, the first two steps in our framework are pseudo labeling and re-sampling. The first step is to use pseudo labeling as an SSL method to predict labels for unlabeled data. The second step involves dividing the dataset into groups based on the protected attribute and the label and then obtaining fair datasets by re-sampling the same number of data points in each group. When unlabeled data is used as training data, it is likely to obtain more under-represented data points from unlabeled data to reduce representation discrimination, and thus to make a little compromise between fairness and accuracy. The issue of noise induced by (incorrectly) predicting labels for unlabeled data is addressed by the third step in the framework: ensemble learning. Predicting unlabeled data will induce some noise in the labels of unlabeled data. Ensemble learning helps to reduce label noise and the variance of the training model, and to produce more accurate final predictions.

In summary, the contributions of this chapter are listed below.

- First, we use unlabeled data to reduce representation discrimination, and thus achieve a better trade-off between accuracy and discrimination.
- Second, we propose a fairness-enhanced sampling (FS) framework that combines pseudo labeling, re-sampling and ensemble learning for fair SSL in the preprocessing phase.
- Third, we theoretically analyze the sources of discrimination in SSL via bias, variance and noise decomposition, and conduct experiments with both real and synthetic data to validate the effectiveness of our proposed FS framework.

4.2 Background

4.2.1 Notations

Let D_l be a dataset with N_1 data points $D_l = \{(x_i, z_i, y_{l,i})\}_{i=1}^{N_1}$, where $x_i \sim X$ contains the information of k unprotected attributes; z_i denotes the protected attribute (also called sensitive attribute), and $y_{l,i} \in \{0,1\} \sim Y$ denotes the label. Let D_u be a dataset with N_2 data points $D_u = \{(x_i, z_i, y_{u,i})\}_{i=1}^{N_2}$, where $y_{u,i}$ denotes the potential label for the unlabeled data. For ease, assume the protected attribute is binary valued. For example, if the protected attribute is race, the value might be either 'white' ($z = 0$) or 'black' ($z = 1$).

Our objective is to learn a mapping $f(\cdot)$ over a discriminatory dataset D_l and D_u, in which the classification result is independent of protected attributes. Performance is measured by both accuracy and the level of discrimination in the results. The ideal classifier should have a high accuracy without discrimination.

4.2.2 Fairness Metrics

Fairness is often evaluated with respect to protected/unprotected groups of individuals defined by attributes, such as gender or age. Here, we have opted for demographic parity as the fairness metrics in this chapter.

Definition 4.1. (Demographic parity) [15] Demographic parity requires that the probability of a classifier's prediction be independent of any sensitive attributes, where the probability of the predicted positive labels in a group z is defined as follows:

$$\gamma_1(\hat{y}) = Pr(\hat{y} = 1 | z = 1) \tag{4.1}$$

$$\gamma_0(\hat{y}) = Pr(\hat{y} = 1 | z = 0) \tag{4.2}$$

where $\hat{y}$ denotes the predicted label.

Definition 4.2. (Discrimination level) The discrimination level γ in terms of demographic parity can be evaluated by the difference between groups,

$$\Gamma(\hat{y}) = |\gamma_0(\hat{y}) - \gamma_1(\hat{y})| \tag{4.3}$$

4.2.3 Discrimination Sources

Discrimination can exist in every stage of machine learning. Roughly, discrimination sources can be divided into two lines: data discrimination and model discrimination [127]. Our proposed FS method is able to reduce the representation discrimination in the data.

4.2.3.1 Data Discrimination

Data discrimination includes historical discrimination, representation discrimination, and measurement discrimination. Historical discrimination occurs when there is a discrepancy between the world itself and the values or goals in the model to be encoded and propagated. It can stem from cultural stereotypes among people, such as social class, race, nationality, and gender. Representation discrimination occurs when the data used to train the algorithm does not accurately represent the problem space. As a consequence, the model generalizes to fit the majority groups much more than minority groups. Measurement discrimination comes from the way we choose, utilize, and measure specific features. The selected set of features and labels may miss important factors, or bring in a group or input-related noise that causes different performance.

4.2.3.2 Model Discrimination

Model discrimination includes aggregation discrimination, evaluation discrimination, and deployment discrimination. Aggregation discrimination can arise during model construction when different populations are improperly grouped together. In many applications, the groups of interest are heterogeneous, so a single model is unlikely to fit all subgroups. Evaluation discrimination occurs during the model iteration and evaluation. This can happen when a test or external benchmark unequally represents each group in the population. Evaluation discrimination may also occur due to the use of performance metrics that are not appropriate for the way the model is used. Deployment discrimination occurs after the model is deployed when the system is used or interpreted in an inappropriate way.

4.2.4 Bias, Variance and Noise

Following [26], our analysis of discrimination is based on bias, variance and noise decomposition. First, we present the definition of the main prediction. The main prediction for a loss function L and set of training sets D is defined as, $y_m(x,a) = \underset{y'}{\operatorname{argmin}} E_D[L(y,y')|X = x, A = a]$, where y is the true value; y' is the predicted label with the minimum average loss relative to all the predictions. The expectation is taken with respect to the training sets in D.

Definition 4.3. (Bias,variance and noise) Following [42], the bias B, variance V and noise N at a point (x,z) are defined as,

$$B(\hat{y},x,z) = L(y^*(x,z), y_m(x,z)) \tag{4.4}$$

$$V(\hat{y},x,z) = E_D[L(y_m(x,z), \hat{y}(x,z)] \tag{4.5}$$

$$N(\hat{y},x,z) = E_D[L(y^*(x,z), y(x,z)] \tag{4.6}$$

where y^* is the optimal prediction that achieves the smallest expected error.

The bias-variance-noise decomposition is suitable for discrimination analysis because the loss is related to the misclassification rate. Loss function can be decomposed into the false positive rate and false negative rate. When these rates are known, true positive rate and true negative rate can also be calculated. This means that many fairness metrics, such as demographic parity and equal opportunity, might be explained via bias, variance and noise decomposition.

4.3 The Proposed Method

4.3.1 Overview of the Fairness-enhanced Sampling Framework

Figure 4.1 shows the general description of the fairness-enhanced sampling framework in the pre-precessing phase. The framework consists of three steps: 1) pseudo labeling, 2) re-sampling and 3) fair ensemble learning. The first step is to predict labels for unlabeled data as more data points in the protected group are likely to be found in unlabeled data. The second step is to construct new datasets that is able to represent all groups equally when the datasets are used for training. In this way, representation

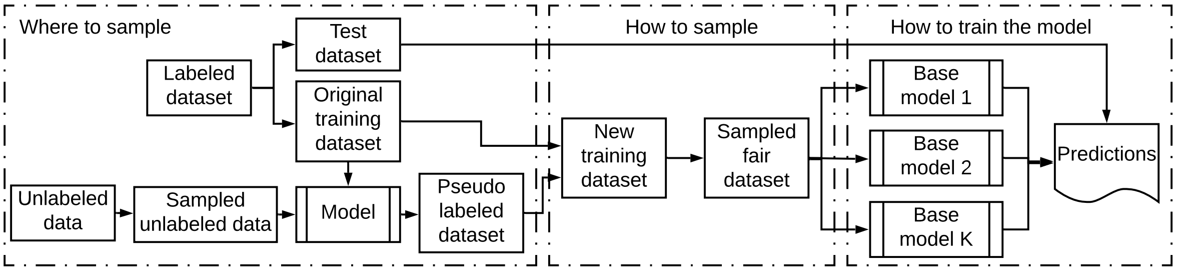


Figure 4.1: The three phases of the fairness-enhanced sampling framework: 1) where to sample, 2) how to sample and 3) how to train the model. Step 1 is to generate a new training dataset which consists of the original dataset and the pseudo labeled dataset. Step 2 is to construct multiple fair datasets through re-sampling. Step 3 is to train a model with each of the fair datasets through ensemble learning to produce the final predictions.

discrimination can be removed from training datasets. The third step is to train multiple base models based on multiple fair datasets and final predicted results are obtained from multiple base models. Ensemble learning is able to reduce the label noise that is induced via pseudo labeling, and the model variance. Each of these steps is discussed in more detail in the following.

4.3.2 Where to Sample

The goal of this step is to use a labeled dataset and part of an unlabeled dataset to construct a new training dataset, as shown in Figure 4.1. Suppose we have a labeled dataset D_l and a large unlabeled dataset D_u. First, we use the labeled dataset and part of the unlabeled dataset to generate a new training dataset. With a sample ratio of ρ, we take random samples from the unlabeled dataset D_u and form sampled unlabeled datasets D_{su}. Then we use pseudo labeling to predict the labels for unlabeled data as if they were true labels. Pseudo labeling is a simple and efficient method to implement SSL [85]. The procedure, as shown in **Algorithm 1**, is as follows.

1) Set a split rate $s \in (0,1)$ and split the labeled dataset into training and test dataset, denoted as the original training dataset and test dataset. 2) Select a learning model and, train the model on the original training dataset to produce a trained model. 3) Use the trained model on D_{su} to predict the output (or pseudo label), and the pseudo labeled dataset is obtained. We do not know if these predictions are correct, but we

now have predicted labels, which is our goal in this step. 4) Concatenate the original training dataset and pseudo labeled dataset to form a new training dataset D_{new}. Pseudo-labeling is an easy-to-implement and efficient semi-supervised learning method

Algorithm 3 Pseudo labeling

```
Input: Labled dataset $D_l$, unlabeled dataset $D_u$, split rate $s$, sample ratio $\rho$
  Split $D_l$ into the training dataset and the test dataset;
  Sample $D_{su}$ from $D_u$
  Select a learning model and train the model on the training dataset;
  Obtain the trained model
  Use the trained model to predict labels for $D_{su}$;
  Combine the original training dataset and the pseudo labeled dataset to create $D_{new}$;
Output: New training dataset $D_{new}$
```

and, by the above method, can take advantage of unlabeled data to both: a) increase the size of the training set; and b) create more data samples representing minority groups to produce fairer training sets. Moreover, the learning model can be any models, such as logistic regression, neural networks, etc.

4.3.3 How to Sample

In this step, the goal is to sample multiple fair datasets from the new training datasets to ensure fair learning. The rationale for this method is that, since the classifier is trained on non-discriminatory data, its prediction may also be non-discriminatory [73]. For simplicity, this analysis covers a binary classification task with one protected attribute, and applies demographic parity as the fairness metric. Our method can certainly be applied to cases with multiple sensitive attributes, subjected to the fairness metrics.

Based on this setup, the dataset is divided into four groups according to the protected attribute and labeled-values: 1) Protected group with positive labels (G_{PP}), 2) Unprotected group with positive labels (G_{UP}), 3) Protected group with negative labels (G_{PN}), and 4) Unprotected group with negative labels (G_{UN}). These divided groups can be denoted as follows,

$$G_{PP} = \{x \in D | z = 1, y = 1\} \tag{4.7}$$

$$G_{UP} = \{x \in D | z = 0, y = 1\} \tag{4.8}$$

$$G_{PN} = \{x \in D | z = 1, y = 0\} \tag{4.9}$$

$$G_{UN} = \{x \in D | z = 0, y = 0\} \tag{4.10}$$

where $y = 1$ denotes the positive class and $y = 0$ denotes the negative class. $z = 1$ denotes that the data point is in the protected group and $z = 0$ denotes that the data point is in the unprotected group. To ensure fair learning in the pre-processing phase, the number of data points in the training set for each group should be the same, otherwise, the model will fall prey to data discrimination. In the case of discrimination, the size of each group is different. Our aim is to adjust the data points by sampling to reach the same size in each group.

Algorithm 2 describes the process of how to obtain multiple fair datasets, and the procedure is as follows: First, we compute the size of the groups G_{PP}, G_{UP}, G_{PN}, G_{UN}. The sample size is denoted as $|G|$, which means that the number of $|G|$ data points will be sampled from each group. Here, there are two cases: 1) When $|G_i| \geq |G|$, $|G|$ data points are sampled randomly from the group G_i. 2) When $|G_i| < |G|$, $|G|$ data points are oversampled from the group G_i. Then we can obtain the fair dataset D_{sf} which consists of the number of data points equally for each of the four groups. Repeating this procedure K times produces K fair datasets with some commonalities and some differences due to the random sampling, which is desirable for ensemble learning. The next step is to learn from these multiple fair datasets to achieve more accurate and less discriminatory results.

4.3.4 How to Train the Model

In this step, the goal is to achieve more accurate and less discriminatory training results on multiple fair datasets D_{sf}. After obtaining multiple D_{sf}, we choose a learning model to train multiple D_{sf} and apply ensemble learning to combine the learning results. Ensemble learning in machine learning exploits the independence between base models to improve the overall performance. In this case, we use Bagging [14] to combine the

Algorithm 4 Fair re-sampling

```
Input: New training dataset $D_{new}$, sensitive attribute $z$, sample times $K$, sample size $|G|$, sample ratio $\rho$
  Divide the dataset into four groups $G_{PP}$, $G_{PN}$, $G_{UP}$, $G_{UN}$
  Calculate the size of all groups $|G_i|$
  for $k \in K$ do
    if $|G_i| \geq |G|$ then
      Sample randomly the number of $|G|$ data points from the group $i$
    end if
    if $|G_i| \leq |G|$ then
      Oversample the number of $|G|$ data points from the group $i$
    end if
    Obtain fair datasets $D_{sf,i}$
  end for
  Obtain multiple fair datasets $D_{sf,1}, D_{sf,2}, ..., D_{sf,K}$
Output: Fair datasets $D_{sf}$
```

decisions from multiple base models learned on multiple fair datasets to improve the accuracy and decrease the discrimination.

Algorithm 3 describes the fair ensemble learning. With the new training dataset D_{new} from **Algorithm 1** and fair datasets $D_{sf,1}, D_{sf,2}, ..., D_{sf,K}$ from **Algorithm 2**, train each fair dataset on its own model $f_k(D_{sf,k})$ in parallel. The final model will average the outputs based on the aggregation of predictions from all base models. The predictions obtained from most base models are predicted as final predictions, which are presented as,

$$f(\cdot) = argmax_{y \in Y} \sum_{k=1}^{K} I(y = f_k(D_{sf,k})) \tag{4.11}$$

where $I(\cdot)$ is the indicator function, and K is the ensemble size, i.e., the number of fair datasets.

Having some diversity across the datasets is crucial for ensemble learning. In our approach, the randomness of the fair datasets reflects in two places: 1) randomly sampling the unlabeled dataset D_u, and subsequently, the pseudo labeled dataset process in **Algorithm 1**; and 2) randomly sampling $|G|$ data points for all groups from D_{new} when constructing each fair dataset.

With ensemble learning, the discrimination level is determined by final predictions.

We redefine the discrimination level in ensemble learning as $\gamma_{En} = |Pr(f(\cdot) = 1|z = 1) - Pr(f(\cdot) = 1|z = 0)|$. Overall, a combination of multiple base models helps to decrease discrimination resulting from variance and noise and is able to give a more reliable prediction than a single model.

Algorithm 5 Fair ensemble learning

```
Input:Dataset, sample times $K$, sample size $|G|$, split rate $s$, sample ratio $\rho$
  Execute Algorithm 1 to obtain the new training dataset $D_{new}$
  for $k \in K$ do
    Execute Algorithm 2 to obtain the fair dataset $D_{sf,k}$
    Train the selected model on the fair dataset $D_{sf,k}$ and obtain the base model $f_k(\cdot)$
  end for
  Make predictions using the final model with ensemble size $K$ in Eq.(12)
Output: Accuracy $Acc$, Discrimination level $\gamma$
```

4.3.5 Discussion

In reviewing the complete framework, there are several benefits to this approach, which are worth highlighting.

- Many semi-supervised learning methods can be used to predict labels for unlabeled data, such as graph-based learning and transductive support vector machines [155]. We choose pseudo labeling because it is a commonly used semi-supervised learning technique, which is efficient and easy to implement.
- The proposed FS framework only removes representation discrimination. However, it is likely that many types of discrimination exist in machine learning, such as historical discrimination, and measurement discrimination. Other discrimination can be removed by in-processing or post-processing methods, based on our proposed FS framework.

4.4 Discrimination Analysis

Following [26], we analyze the fairness of the predictive model via bias, variance, and noise decomposition. The source of discrimination can be decoupled as discrimination in bias $B_z(\hat{y})$, discrimination in variance $V_z(\hat{y})$ and discrimination in noise N_z. The expected discrimination level $\Gamma(\hat{y})$ of a classifier f learned from a set of training set D is defined as, $\bar{\Gamma}(\hat{y}) = |E_D[\Gamma_0(\hat{y}) - \Gamma_1(\hat{y})]|$.

Lemma 4.1. *The discrimination with regard to group $z \in Z$ is defined as,*

$$\gamma_z(\hat{y}) = \bar{B_z}(\hat{y}) + \bar{V_z}(\hat{y}) + \bar{N_z} \tag{4.12}$$

When the protected group and unprotected group is given, the discrimination level is calculated as,

$$\bar{\Gamma} = |(\bar{B_0} - \bar{B_1}) + (\bar{V_0} - \bar{V_1}) + (\bar{N_0} - \bar{N_1})| \tag{4.13}$$

and each of the component of Eq. (39) are calculated as,

$$\bar{B_z}(\hat{y}) = E_D[B(y_m, x, z)|Z = z] \tag{4.14}$$

$$\bar{V_z}(\hat{y}) = E_D[c_v(x, z)V(y_m, x, z)|Z = z] \tag{4.15}$$

$$\bar{N_z} = E_D[c_n(x, z)L(y^*(x, z), y)|Z = z] \tag{4.16}$$

where $c_v(x,z)$ and $c_n(x,z)$ are parameters related to the loss function. For more details, see the proof in [26].

Lemma 4.2. *The discrimination learning curve $\bar{\Gamma}(\hat{y}, k) := |\bar{\gamma_0}(\hat{y}, k) - \bar{\gamma_1}(\hat{y}, k)|$ is asymptotic and behaves as an inverse power law curve, where k is the size of the training set [26].*

Theorem 4.1. *Unlabeled data is able to reduce discrimination with the proposed FS framework, if* $(|\bar{V_z}(\hat{y})_{sl}| - |\bar{V_a}(\hat{y})_{ssl}|) - \bar{N}_{z,p} \geq 0$.

Proof. To prove the above theorem, we shall prove that the discrimination level in SSL $\bar{\Gamma}_{ssl}$ is lower than the discrimination level in supervised learning $\bar{\Gamma}_{sl}$. In the following, we will analyze the discrimination in SSL in terms of *discrimination in bias $\bar{B_z}(f)_{ssl}$, discrimination in variance $\bar{V_z}(\hat{y})_{ssl}$, and discrimination in noise $\bar{N}_{z,ssl}$.*

Discrimination in Bias Bias measures the fitting ability of the algorithm itself, and describe accuracy of the model. Hence, bias in discrimination $\bar{B_z}(\hat{y}) = E_D[B(y_m, x, z)|Z = z]$ only depends on the model. When the same model is trained on the original training dataset and new training dataset, discrimination in bias is the same in supervised learning and SSL, which can be expressed as $|\bar{B}(\hat{y})_{sl}| - |\bar{B}(\hat{y})_{ssl}| = 0$.

Discrimination in Variance Discrimination in variance $\bar{V}_z(\hat{y})$ can be reduced with extra unlabeled data in the training dataset. **Lemma 2** states that the discrimination level $\bar{\Gamma}(\hat{y}, n)$ decreases with the increasing size of training data n. In our proposed FS framework, unlabeled data is pseudo-labeled, and the new training dataset consists of the original training dataset and the pseudo labeled dataset. The size of the new training dataset can be guaranteed to be larger than the size of the original training by adjusting the sampling size. Also, using Bagging to combine all the base models to obtain the final predictions helps to construct the aggregate model with a lower variance, thus reducing the discrimination in variance $\bar{B}_z$. Hence, we conclude that $|\bar{V}_z(\hat{y})_{ssl}| - |\bar{V}_z(\hat{y})_{sl}| \leq 0$.

Discrimination in Noise Unlabeled data introduces more discrimination in noise because pseudo labeling contains discrimination from the trained model. Thus, noisy labels from pseudo labeling in the unprotected group is more than that in the protected group. We divide the discrimination in noise in SSL into discrimination in noise in labeled data $\bar{N}_{z,l}$ and discrimination in noise in pseudo labeled data $\bar{N}_{z,p}$, which is expressed as,

$$\bar{N}_{z,ssl} = \bar{N}_{z,l} + \bar{N}_{z,p} \tag{4.17}$$

Discrimination in noise in labeled data $\bar{N}_{z,l}$ is the same as the discrimination in noise in supervised learning $\bar{N}_{z,sl}$. Then we analyze the discrimination in noise due to pseudo labeled data $\bar{N}_{z,p}$, including four mislabeled cases,

$$\bar{N}_{y=0,z=0} = E_{D_{un}}[\hat{y}_p^* = 1 | y = 0, z = 0] \tag{4.18}$$

$$\bar{N}_{y=0,z=1} = E_{D_{un}}[\hat{y}_p^* = 1 | y = 0, z = 1] \tag{4.19}$$

$$\bar{N}_{y=1,z=0} = E_{D_{un}}[\hat{y}_p^* = 0 | y = 1, z = 0] \tag{4.20}$$

$$\bar{N}_{y=1,z=1} = E_{D_{un}}[\hat{y}_p^* = 0 | y = 1, z = 1] \tag{4.21}$$

where $\hat{y}_p^*$ is the optimal predicted label of unlabeled data via pseudo labeling. The noise in the protected group is $\bar{N}_{1,p} = \bar{N}_{y=0,z=1} + \bar{N}_{y=1,z=1}$ and the noise in the unprotected group is $\bar{N}_{0,p} = \bar{N}_{y=0,z=0} + \bar{N}_{y=1,z=0}$. The model contains discrimination because the model is trained on a dataset without any fairness guarantees, and thus the model will bring discrimination in pseudo labeling. In this way, discrimination in noise in pseudo labeled data $\bar{N}_{z,p}$ can be measured as,

$$\bar{N}_{z,p} = |\bar{N}_{1,p} - \bar{N}_{0,p}| \tag{4.22}$$

To relieve the noise from pseudo labeling, we use Bagging - a robust model that is resilient to class label noise since the errors incurred by the noise can be compensated by the combined predictions of other learners.

Based on the analysis above, we conclude that when $|\bar{V}_z(\hat{Y})_{ssl} - \Delta\bar{V}_z(\hat{Y})_{sl}| - \bar{N}_{z,p} \geq 0$, unlabeled data is able to reduce discrimination with the proposed FS framework. Unlabeled data do not change discrimination in bias. However, they do reduce discrimination in variance, and they increase discrimination in noise, but bagging reduces discrimination both in variance and discrimination in noise. ■

4.5 Experiment

In this section, we demonstrate our framework by performing experiments on real-world and synthetic datasets. The goal of our experiments is three folds. The first is to show how the framework makes use of unlabeled data to achieve a better trade-off between accuracy and discrimination. The second is to explore the impact of factors, such as ensemble times and sampling size, on the training results. And, third, we show the distinct difference in discrimination level when the model is tested with discrimination test dataset and fair test dataset.

4.5.1 Experiments on Real Data

The aim of real-world datasets is to assess the effectiveness of our method to achieve a better trade-off between accuracy and discrimination with unlabeled data. We also show the benefit of ensemble learning, the impact of the sampling size, and the comparison with other methods.

4.5.1.1 Experimental Setup

Dataset The experiments involve three real-world datasets: the Health dataset [1], the Bank dataset [2], the Adult dataset [3].

[1] https://foreverdata.org/1015/index.html

[2] https://archive.ics.uci.edu/ml/datasets/bank+marketing

[3] http://archive.ics.uci.edu/ml/datasets/Adult

- The target of Health dataset is to predict whether people will spend any day in the hospital. In order to convert the problem into the binary classification task, we simply predict whether people will spend any day in the hospital or not. Here, ’Age’ is the protected attribute and two groups are divided at ≥65 years. After data pre-processing, the dataset contains 10000 records with 132 features.
- The Bank dataset contains a total of 31,208 records with 20 attributes and a binary label, which indicates whether the client has subscribed to a term deposit or not. Again, ’Age’ is the protected attribute.
- The target of Adult dataset is to predict whether people’s income is larger than 50K dollars or not, and we consider "Gender" as the protected attribute. After data pre-processing, the dataset contains 48,842 records with 18 features.

Parameters The protected attribute is excluded from the prediction model during the training to ensure equity across groups. The protected attribute is only used to evaluate the discrimination measurement in the testing phase. In the above of three real-world datasets, data are all labeled. First, we split the whole dataset randomly into two halves: one half is used as labeled dataset, and we remove the labels from the other half to serve as the unlabeled dataset. In the labeled data, we set the split rate $s = 0.8$, which means 80% of the data are used for training and 20% of the data are used for testing. The sample size n_s equals the minimum size of four groups in three datasets.

The final result is an average of 50 results run in the new training datasets. For each run, we generate $K = 200$ fair datasets and construct with $K = 200$ base models to make the final predictions. We use 5-fold cross-validation on the original training dataset and test dataset.

Baseline Given our method is a pre-processing method, we compare it to two other pre-processing methods and the method without any fair process.

- Original (ORI): The original dataset is used for training without fairness guarantees.
- Uniform Sampling (US) [73]: The number of data points in each group is equalized through oversampling and/or undersampling.

- Preferential Sampling (PS) [73]: The number of data points in each group is equalized by taking samples near the borderline data points.

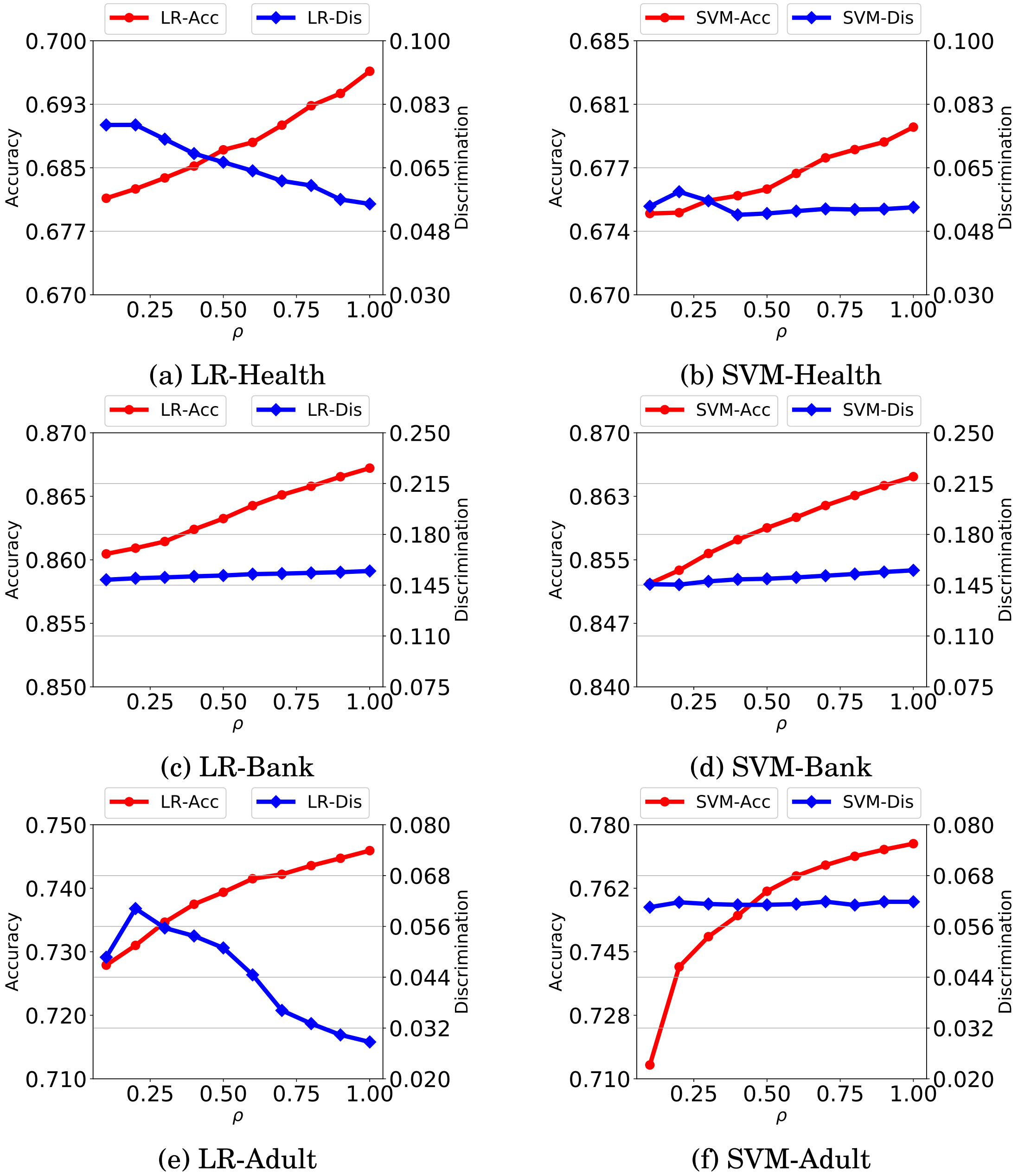
4.5.1.2 Trade-off Between Accuracy and Discrimination

Figure 4.2 shows the accuracy and discrimination level varies given different sample ratios ρ with logistic regression (LR) and support vector machine (SVM) on three datasets. As shown, accuracy generally increases with growing size of unlabeled data. For example, LR has an accuracy of around 0.728 when $\rho = 0.1$ with the Adult dataset, which increases to 0.745 when $\rho = 1$. This indicates that the unlabeled data helps to improve the accuracy to some extent. Also, we note that accuracy relates to the training models and the choice of training models relates to the datasets. The discrimination level has different performances in different training models. For example, with the Adult dataset, the discrimination level initially increases and then steadily decreases till the end in LR. The discrimination level is steady and has a slight increase in SVM. This observation indicates that unlabeled data can help to reduce the discrimination for some models, like LR. Similar to accuracy, the discrimination level relates to the training models and our experiments show that LR is more friendly in discrimination than SVM. The choice of sample ratio depends on the quality of the dataset itself as well as the requirement of the learning task. Accuracy could be improved with unlabeled data, while discrimination level depends on the reduction of discrimination in variance and increase of discrimination in noise that unlabeled data could bring in the training.

4.5.1.3 The Impact of Ensemble Learning

Figure 4.3 shows the impact of ensemble learning on accuracy and discrimination level with LR and SVM on three datasets. In ensemble learning, we sample the percentage of $\rho = 1$ unlabeled data from the unlabeled dataset, and generate the new training dataset. With LR, the accuracy typically increases and then steadies till the end, whereas, with SVM accuracy fluctuates before steadying at some lower, equal or higher rate. This is because the errors in variance and noise reduction as the ensemble size increases.

In terms of discrimination levels, both methods show fluctuations at first before stabilizing on all three datasets. The changes in discrimination levels have no obvious correlations to accuracy prior to convergence. This is reasonable because training results



(a) LR-Health

(b) SVM-Health

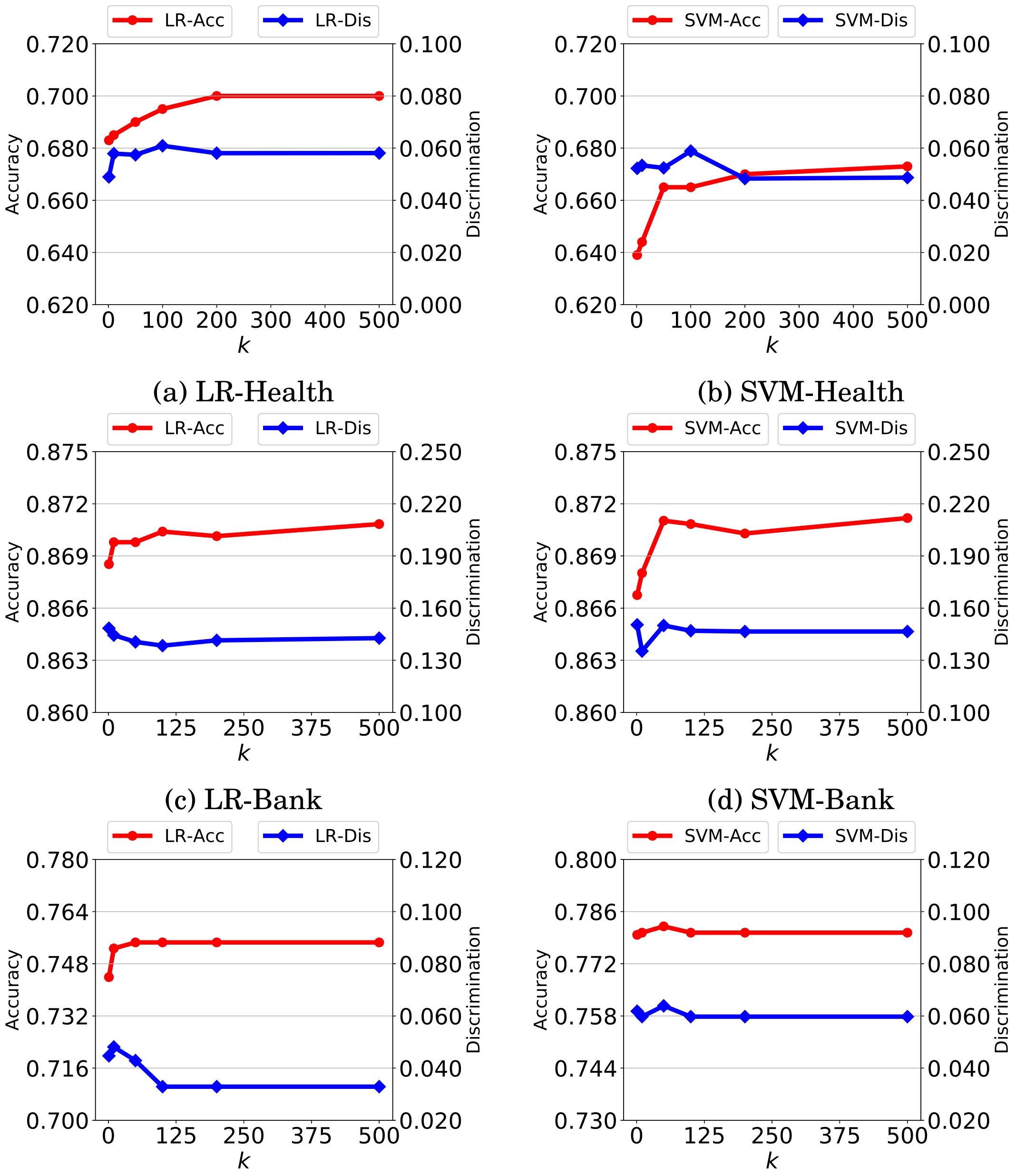
(c) LR-Bank

(d) SVM-Bank

(e) LR-Adult

(f) SVM-Adult

Figure 4.2: The trade-off between accuracy (Red) and discrimination level (Blue). (a) LR in Health dataset; (b) SVM in Health dataset; (c) LR in Bank dataset; (d) SVM in Bank dataset; (e) LR in Adult dataset; (f) SVM in Adult dataset. The X-axis is the sample ratio ρ, which denotes that the percentage of ρ unlabeled data are sampled from the unlabeled dataset and then pseudo labeled for training.



(a) LR-Health

(b) SVM-Health

(c) LR-Bank

(d) SVM-Bank

(e) LR-Adult

(f) SVM-Adult

Figure 4.3: The impact of ensemble learning on the accuracy (Red) and discrimination level (Blue) on (a) LR in Health dataset; (b) SVM in Health dataset; (c) LR in Bank dataset; (d) SVM in Bank dataset; (e) LR in Adult dataset; (f) SVM in Adult dataset. Initially, there is not an obvious link between accuracy and discrimination level. However, as the ensemble size grows, the accuracy and discrimination levels begin to converge. Each point is an average of 50 times.

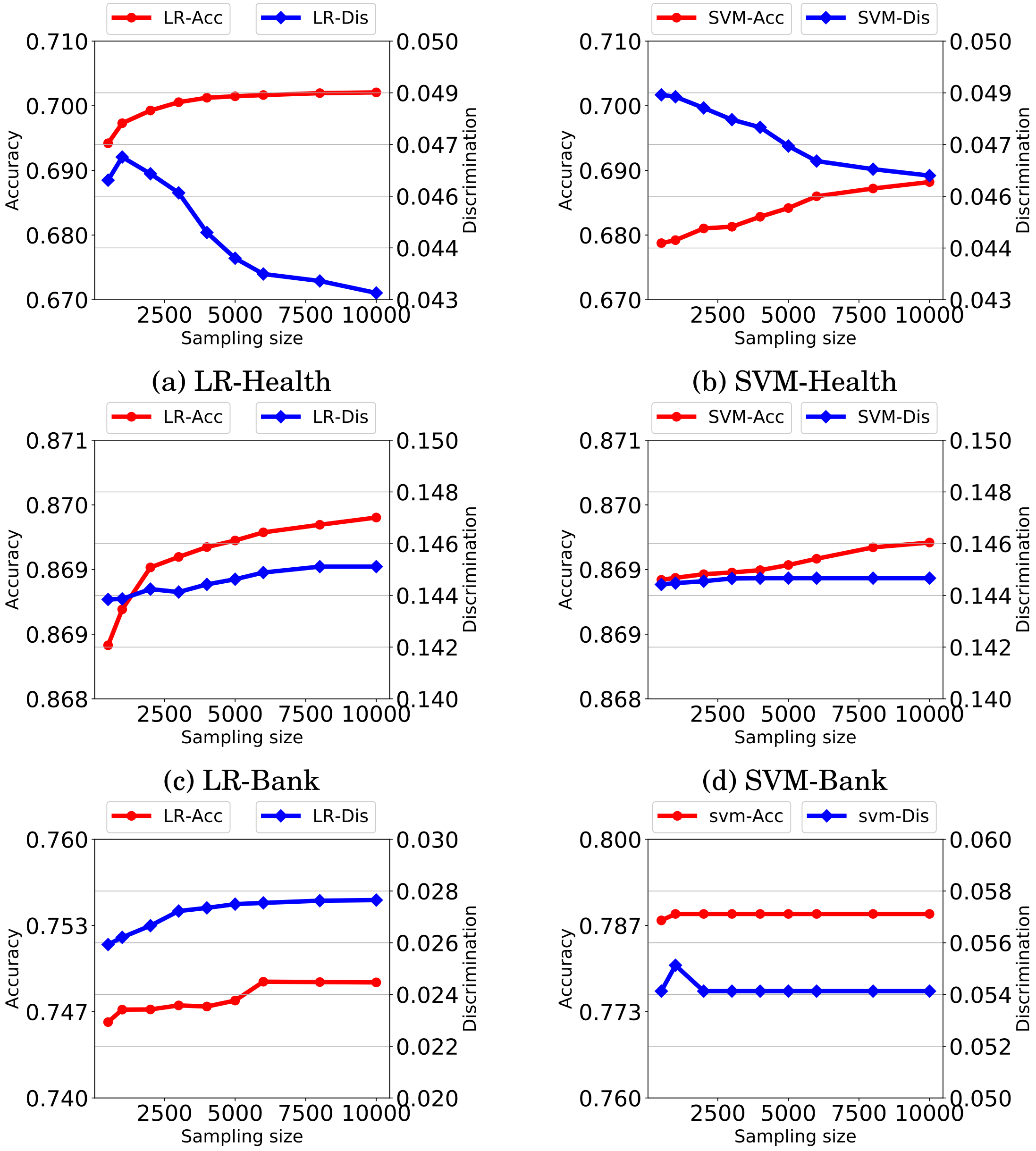
having the same accuracy does not mean the same discrimination level. Also, without a sufficient ensemble size, training on fair datasets will introduce some variance and noise to the final result. Overall, an ample ensemble size helps to improve accuracy and decrease discrimination. The appropriate ensemble size is $K = 200$ or so. This is because accuracy increases and discrimination fluctuates before $K = 200$, and broadly accuracy and discrimination become steady after $K = 200$ for three datasets.

4.5.1.4 The Impact of Sample Size

Figure 4.4 shows the impact of sample size on accuracy and discrimination level with LR and SVM on three datasets. Overall, it is observed that accuracy increases quickly in the early stages and then becomes stable as the sample size grows. This is because more data help to improve the generalization ability, but extra data do not help when the amount of data is enough to fit the model. Unlike accuracy, discrimination level depends on the amount of label noise that unlabeled data may bring when the sample size increases. For example, discrimination decreases in the Health dataset and increases a litter in the Bank dataset. This means that, with an increasing of sample size, little label noise is brought into the Health dataset, and consequently discrimination level decreases. Also, it is note that LR is more sensitive to sample size than SVM. The choice of sample size depends on the quality of the dataset and the training task requirement. Generally, a larger sample size can improve accuracy, reduce discrimination in bias and increase discrimination in noise.

4.5.1.5 Comparison with other methods

Figure 4.5 shows the results from a comparison of our proposed FS method with and the other three schemes in terms of the accuracy and discrimination level on the three datasets. The training dataset of other methods is the original training dataset and the training dataset of our method is the new training dataset that consists of the original training dataset and pseudo labeled dataset (ρ=1). The test dataset is the same. The results show that our method is able to push the discrimination to very low values while achieving a fairly high accuracy compared with other schemes. Specifically, in the Adult dataset, the discrimination level under LR is around 0.215 with the original method and around 0.022 with the preferential sampling method, and the proposed FS method



(a) LR-Health

(b) SVM-Health

(c) LR-Bank

(d) SVM-Bank

(e) LR-Adult

(f) SVM-Adult

Figure 4.4: The impact of sample size on accuracy (Red) and discrimination level (Blue) on (a) LR in Health dataset; (b) SVM in Health dataset; (c) LR in Bank dataset; (d) SVM in Bank dataset; (e) LR in Adult dataset; (f) SVM in Adult dataset. An increasing in the sampling size leads to an increase in accuracy and may help to reduce discrimination level.

can decrease discrimination to 0.019 with better accuracy than the preferred sampling method. This indicates that the proposed FS method is able to reduce discrimination better than other methods.

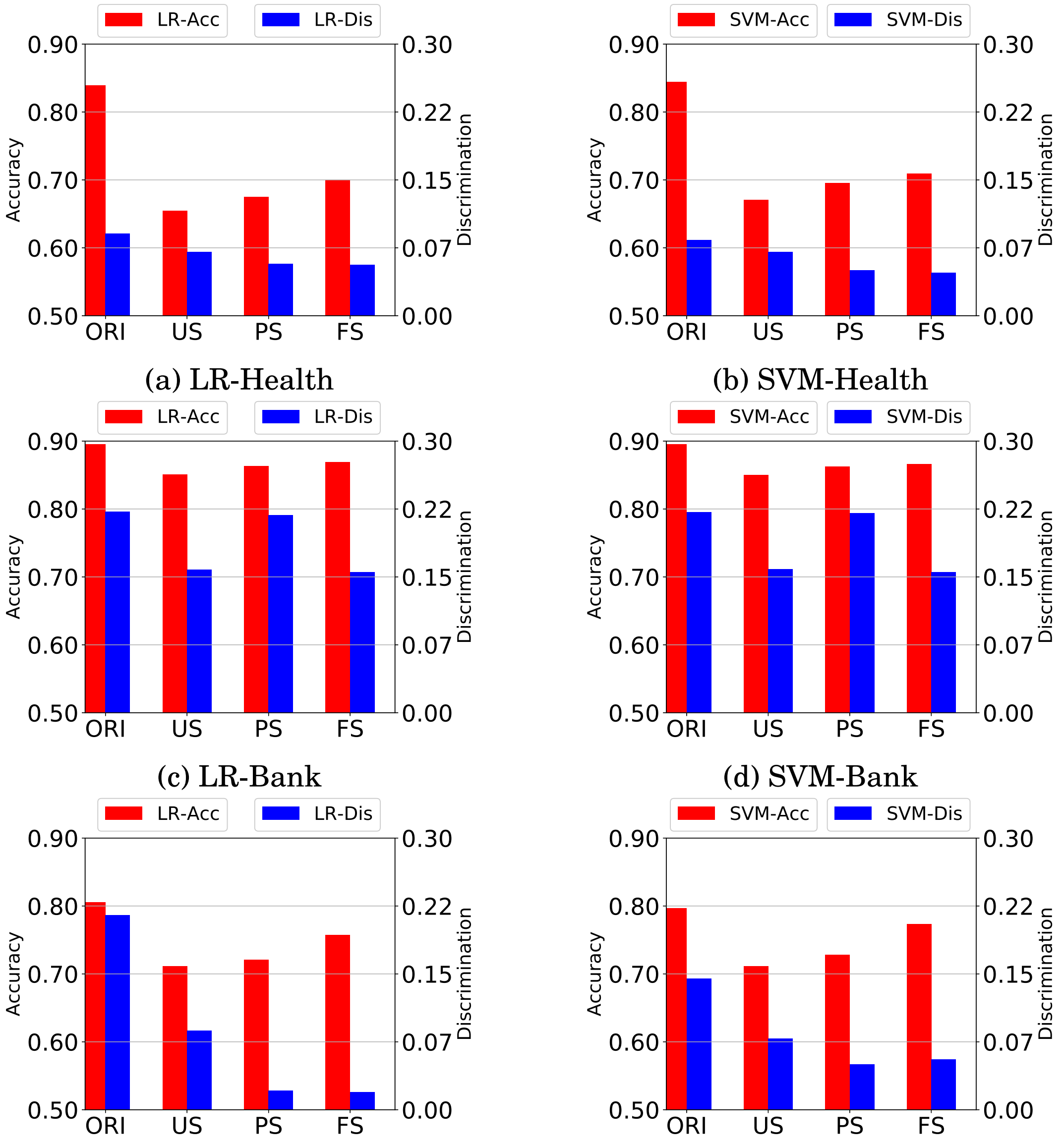
4.5.2 Experiments on Synthetic Data

We first describe how to generate synthetic datasets and the goal of synthetic datasets is to show the effectiveness of our method in the discriminatory test dataset and fair test dataset. Here, the discriminatory test dataset refers to the test dataset whose data points are not equally presented in each group, and the fair test dataset refers to the test dataset whose data points are equally presented in each group. We show the distinct difference in discrimination on two types of test datasets.

4.5.2.1 Synthetic Data Setup

We generate 22,000 binary class labels and a protected attribute a with a uniform random distribution, and assign a 2-dimensional feature vector to each label by drawing samples from two different Gaussian distributions: $p(x|y=1) = N([2;2],[5,1;1,5])$ and $p(x|y=-1) = N([-2;-2],[10,1;1,3])$. The size of each group in the synthetic dataset is roughly the same. Then we randomly sample 2,000 data points from the synthetic dataset as a fair test dataset, and split the remaining dataset randomly into two halves: one half is to be used as the labeled dataset and the other half with labels removed to serve as the unlabeled dataset.

Note that the synthetic dataset is a fair dataset, and the discriminatory dataset is generated by calibrating data points in the group G_{PP} based on the synthetic dataset. Discriminatory dataset 1 (DA 1) is generated by sampling 2,000 data points randomly in the group G_{PP} and data points do not change in other groups. Discriminatory dataset 2 (DA 2) is generated by sampling 3,000 data points randomly in the group G_{PP} and data points do not change in other groups. In each discriminatory dataset, we sample 2,000 data points as the discriminatory test dataset and the remaining as the training dataset.



(a) LR-Health

(b) SVM-Health

(c) LR-Bank

(d) SVM-Bank

(e) LR-Adult

(f) SVM-Adult

Figure 4.5: Comparison with original scheme (ORI), uniform sampling (US) and preferential sample (PS) with (a) LR in Health dataset; (b) SVM in Health dataset; (c) LR in Bank dataset; (d) SVM in Bank dataset; (e) LR in Adult dataset; (f) SVM in Adult dataset. With the fairness-enhanced sampling method (FS), discrimination decreases without much cost of accuracy or accuracy increases without much cost of discrimination.

	Test with discriminatory test dataset						
	Method	Acc	Dis	G_{PP}	G_{UP}	G_{PN}	G_{UN}
LR	DA 1 (ORI)	0.8815	0.2705	183	586	626	605
	DA 2 (ORI)	0.8875	0.3642	104	628	642	626
	DA 1 (FS)	0.8825	0.2076	232	537	627	604
	DA 2 (FS)	0.8730	0.2890	159	573	642	626
SVM	DA 1 (ORI)	0.8825	0.2664	188	581	629	602
	DA2 (ORI)	0.8880	0.3724	102	630	649	619
	DA 1 (FS)	0.8825	0.2097	231	538	628	603
	DA 2 (FS)	0.8745	0.3130	149	583	655	476

Table 4.1: Two discriminatory datasets tested on the discriminatory test dataset in ORI method and the proposed fairness-enhanced method (FS) with LR and SVM. We show accuracy (Acc), discrimination level (Dis) and the number of data points of each group in the discriminatory test dataset after classification.

4.5.2.2 Synthetic Data Tested with Discriminatory and Fair Datasets

Table 4.1 shows that our method is able to reduce discrimination level when training datasets have different discrimination levels. For example, more data points are classified into the Protected group with positive labels G_{PP} after implementing our method, and discrimination level of DA 1 reduces from 0.2705 to 0.2076 in LR. It is also note that accuracy does not decrease much with the proposed FS method. For example, accuracy of DA 2 reduces from 0.8825 to 0.8730 in LR.

	Test with fair test dataset						
	Method	Acc	Dis	G_{PP}	G_{UP}	G_{PN}	G_{UN}
LR	DA1 (ORI)	0.8701	0.0484	438	556	492	514
	DA 2 (ORI)	0.8535	0.1018	376	618	483	523
	DA 1 (FS)	0.8790	0.0161	474	520	496	510
	DA 2 (FS)	0.8810	0.0062	471	523	483	523
SVM	DA1 (ORI)	0.8700	0.0483	441	553	495	511
	DA 2 (ORI)	0.8525	0.1118	372	622	489	517
	DA1 (FS)	0.8790	0.0168	474	520	496	510
	DA 2 (FS)	0.8775	0.0272	460	534	493	513

Table 4.2: Two discriminatory datasets tested on the fair test dataset in ORI method and the proposed fairness-enhanced method (FS) with LR and SVM. We show accuracy (Acc), discrimination level (Dis) and the number of data points of each group in the fair test dataset after classification.

We test the biased datasets with the proposed FS method on the fair test dataset with LR and SVM, and results are shown in Table 4.2. With the proposed FS method, discrimination level decreases and accuracy increases. More specifically, discrimination level decreases from 0.1018 to 0.0062 and accuracy increases from 0.8535 to 0.8810 in the DA 2. Discrimination level with the discriminatory test dataset is much higher than with the fair test dataset. We attribute this to the evaluation bias. Discriminatory dataset and discriminatory test data have the same data distribution, and thus the size of each group in the discriminatory test dataset is not equal. Even if the trained classifier is fair, the result may still be unfair. In real-world datasets, test datasets are sampled from the whole datasets and thus can contain evaluation bias.

4.5.3 Discussion and Summery

4.5.3.1 Discussion

We discuss on how the proposed FS framework is able to reduce discrimination in terms of discrimination decomposition into discrimination in bias, variance and noise. Discrimination in bias depends on the model choice. As we observe in the experiments, very broadly, LR can achieve a lower discrimination level than SVM. Discrimination in variance relates to the training data. Unlabeled data help to reduce discrimination in variance by increasing the size of training data. Ensemble learning helps to reduce discrimination in variance by averaging the training results from base models. An appropriate unlabeled data size, sample size and ensemble size in our framework is able to help reduce more discrimination in variance. Discrimination in noise depends on the quality of data. Training with unlabeled data may bring discrimination in noise. However, ensemble learning offsets this effect. When the same model is used, the benefit of unlabeled data in discrimination reduction depends on the impact of unlabeled data on discrimination in variance and discrimination in noise.

4.5.3.2 Summary

From these experiments, we see that the FS framework is able to reduce representation discrimination with a better trade-off between accuracy and discrimination. In the proposed FS framework, discrimination reduction in variance is usually more than the discrimination incurred by label noise. However, all the factors in the framework, model choice, unlabeled data size, ensemble size, sample size - each make their own particular contribution to increasing accuracy while ensuring fair representation.

4.6 Related Work

In recent years, much research on fair machine learning has been undertaken. The following subsections summarize the three main streams of this work.

4.6.1 Pre-processing Methods

Pre-processing methods eliminate the discrimination by adjusting the training data by ways of suppression, reweighing or sampling to obtain fair datasets before training [15, 17, 73]. Also, learning fair intermediate representations in the pre-process phase has received much attention. [147] was the first to open up fair machine learning by learning fair intermediate representations. The basic idea is that mapping the training data to a transformed space where as much useful information as possible is retained, but the dependencies between sensitive attributes and class labels are removed. Many researchers have subsequently studied fair representation learning with different methods, such as adversary learning [52, 94, 99, 124, 148]. These methods are based on using a classifier to predict sensitive attributes as adversarial components. The advantage of pre-precessing methods is that these methods can apply to all algorithms and tasks. Note that pre-processing approaches cannot be employed to eliminate discrimination arising from the algorithm itself.

4.6.2 In-processing Methods

In-processing methods avoid discrimination with fair constraints [75] used regularizer term to penalize discrimination to enforce non-discrimination in the learning objective. [43, 145, 146] designed fairness constraints to achieve fair classification, where the fairness constraint is enforced by weakening the correlation between sensitive attribute and labels. In [4, 36], the constrained optimization problem is formulated as a two-player game and fairness definitions are formalized as linear inequalities. Other recent works have a similar spirit to enforcing fairness by adding constraints to the objective [5, 82]. The advantage of in-processing methods is that the level of fairness and accuracy can be controlled by the threshold of fairness constraints. However, fairness constraints are often irregular and need to be relaxed for optimization, and thus the solution may not be convergent. In addition, individual fairness can also be regarded as in-processing methods [45, 46, 154].

4.6.3 Post-processing Methods

A third approach to achieving fairness is post-processing, where a learned classifier is modified to adjust the decisions to be non-discriminatory for different groups. [60] proposed an approach to use of post-processing to ensure fairness criteria of equal opportunity and equal odds and subsequent work include [79, 91] However, it is not guaranteed to find the most accurate fair classifier [135], and requires test-time access to the protected attribute, which might not be available.

4.6.4 Comparison with Other Work

Existing fair methods focus on supervised and unsupervised learning, and these methods cannot be applied to SSL directly. As far as we know, only [33, 104] considered fair SSL. In [33], data is used to learn the output conditional probability, and unlabeled data is used for calibration in the post-processing phase. This method is to eliminate the aggregation discrimination, while the proposed FS method is to reduce representation discrimination. In [104], the proposed method is built on neural networks for SSL in the in-processing phase, and this method is to reduce measurement discrimination. In [73], representation discrimination is reduced by uniform sampling and preferential sampling, while in some cases not enough data in minority groups can be sampled to generate a fair dataset. Our work makes use of unlabeled data to form fairer datasets and theoretically analyze the discrimination via decomposition in bias, variance and noise. In this chapter, we study the fair SSL based on labeled and unlabeled data in the pre-processing phase and our goal is to use labeled data to reduce representation discrimination, and in turn achieve a better trade-off between accuracy and discrimination.

4.7 Summary

In this chapter, we use unlabeled data to achieve a better trade-off between accuracy and discrimination in the pre-processing phase. To achieve this, we developed a three-pronged strategy, where each component makes an important contribution to decreasing discrimination and/or improving the accuracy of the final predictions. Pseudo labeling in a semi-supervised setting exploits unlabeled data, on the premise that more training data is likely to reduce discrimination. A re-sampling method leads to multiple sampled fair datasets, and training on fairly-sampled will result in a fairly trained model. Lastly, ensemble learning is applied to improve the quality of the final predictions. Theoretical analysis and our experimental results show that our method delivers what it promises - unlabeled data is a viable option to achieve a better trade-off between accuracy and discrimination. Model choice, unlabeled data size, ensemble size and sampling size are factors that affect training results.

CHAPTER 5

FAIRNESS CONSTRAINTS IN SEMI-SUPERVISED LEARNING

In the last chapter, we studied how unlabeled data help with a better trade-off between model accuracy and fairness in the pre-processing phase. In real-world practice, controlling the trade-off between accuracy and fairness is important. However, pre-processing methods are not able to control the level of fairness, while in-processing methods can control the level of fairness by adjusting the fairness constraint. Therefore, we continue to study fairness in semi-supervised learning in this chapter, but from the perspective of in-processing.

5.1 Introduction

Machine learning algorithms, as useful decision-making tools, are widely used in society. These algorithms are often assumed to be paragons of objectivity. However, many studies show that the decisions made by these models can be biased against certain groups of people. For example, Abid et al. observed that large-scale language models capture undesirable racial bias [2] and Vigdor et al. [131] reported gender bias in the credit ranking of Apple card. These events prove that discrimination can arise from machine learning, and one of the most important discrimination sources is data, including data collection (imbalanced training set) and data preparation (biased content in the training set) [127].

In recent years, many fairness metrics have been proposed to define what is fairness in machine learning. Popular fairness metrics include statistical fairness [31, 146], individual fairness [45, 46, 72] and casual fairness [83, 136]. Meanwhile, a great many algorithms have been developed to address fairness issues for both supervised learning settings [60, 124, 146] and unsupervised settings [8, 28, 30, 116]. Generally, these studies have focused on two key issues: how to formalize the concept of fairness in the context of machine learning tasks, and how to design efficient algorithms that strike a desirable trade-off between accuracy and fairness. What is lacking is the research that considers semi-supervised learning (SSL) scenarios.

In real-world machine learning tasks, big data used for training is necessary and is often a combination of labeled and unlabeled data. Therefore, fair SSL is a vital area of development. Like the other learning settings, achieving a balance between accuracy and fairness is a key issue. According to [26], increasing the size of the training set can create a better trade-off. This finding sparked an idea over whether the trade-off might be improved via unlabeled data. Unlabeled data is abundant in the era of big data and, if it could be used as training data, we may be able to make a better compromise between fairness and accuracy. To achieve this goal, two challenges are ahead of us: 1) how to achieve fair learning from both labeled and unlabeled data; and 2) how to give labels for unlabeled data to ensure that the learning is towards a fair direction.

To solve these challenges, we propose a framework of fair SSL that can support multiple classifiers and fairness metrics. The framework is formulated as an optimization problem, where the objective function includes a loss for both the classifier and label propagation, and fairness constraints over labeled and unlabeled data. Classifier loss is to optimize the accuracy of training results; label propagation loss is to optimize the label predictions on unlabeled data; the fairness constraint is to adjust the fairness level as desirable. The optimization includes two steps. In the first step, fairness constraints enforce weights update towards a fair direction. This step can be solved by a convex problem and convex-concave programming when disparate impact and disparate mistreatment are used as fairness metrics respectively. In the second step, updated weights further direct labels assigned to unlabeled data in a fair direction by label propagation. Labels for unlabeled data can be calculated in a closed form. In this way, labeled and unlabeled data are used to achieve a better trade-off between accuracy and fairness. With this strategy, we can control the level of discrimination in the model and, therefore, provide a machine learning framework that offers fair SSL. Our approach incorporates a wide range of fairness definitions such as disparate impact and disparate mistreatment, which is guaranteed to yield an accurate fair classifier.

With the aim of achieving fair SSL, the contributions of this chapter are three-fold.

- First, we propose a framework that is able to achieve fair SSL that supports multiple classifiers and fairness metrics of disparate impact and disparate mistreatment. This framework enables the use of unlabeled data to achieve a better trade-off between fairness and accuracy.

- Second, we propose algorithms to solve optimization problems when disparate impact and disparate mistreatment are integrated as fairness metrics in the graph-based regularization.

- Third, we consider different cases of fairness constraints on labeled and unlabeled data, and analyze the impacts of these constraints on the training results. This helps us how to control the fairness level in practice.

- Forth, we theoretically analyze the sources of discrimination in SSL, and conduct extensive experiments to validate the effectiveness of our proposed method.

5.2 Preliminaries

5.2.1 Notations

Let $X = \{x_1, ..., x_k\}^T \in \mathbb{R}^{N \times (k+1)}$ denote the training data matrix, where N is the number of data point and k is the number of unprotected attributes; $\mathbf{z} = \{z_1, ...z_N\} \in \{0,1\}^N$ denotes the protected attribute, e.g., gender or race. Labeled dataset is denoted as $D_l = \{x_i, z_i, y_{l,i}\}_{i=1}^{N_1}$ with N_1 data points, and $\mathbf{y_l} = \{y_{l,1}, ..., y_{l,N_1}\}^T \in \{0,1\}^{N_1}$ is the label for the labeled dataset. Unlabeled dataset is denoted as $D_u = \{x_i, z_i\}_{i=1}^{N_2}$ with N_2 data points, and $\mathbf{y_u} = \{y_{u,1}, ..., y_{u,N_2}\}^T \in \{0,1\}^{N_2}$ is the predicted labeled for the unlabeled dataset.

Given the whole dataset, an adjacent matrix is denoted as $A = \theta_{ij} \in \mathbb{R}^{N \times N}, \forall i, j \in 1, ..., N, (N = N_1 + N_2)$, where θ_{ij} is the weight to evaluate the relationship of two data points. The degree matrix D_{DM} is constructed as a diagonal matrix whose i-th diagonal element is $d_{ii} = \sum_{j=1}^{N} \theta_{ij}$. We use L to denote Laplacian matrix, calculated as $L = D_{DM} - A$. Our objective is to learn a classification model $f(\cdot)$ with the model parameters $\mathbf{w}$ (or W) and $\mathbf{y_u}$ over discriminatory datasets D_l and D_u that delivers high accuracy with low discrimination.

5.2.2 Graph-base Regularization

In graph-based regularization, the goal is searching for a function f on the graph. f has to satisfy two criteria simultaneously: 1) it should be as close to the given labels as possible, and 2) it should be smooth on the entire constructed graph. Graph stores the geometric structure in the data (such as similarity or proximity) and uses this structure as a regularizer to infer labels of unlabeled data. Generally, the graph-based regularization methods adopt the following objective function,

$$\mathscr{L} = \mathscr{L}_C + \alpha \mathscr{L}_R \tag{5.1}$$

where $\mathscr{L}_C$ is the classification loss; α is a balancing parameter; $\mathscr{L}_R$ is a graph-based regularizer. Different methods can have different variants of the regularizer. In this work, we consider Laplacian regularizer as it is the most common used regularizer, which is calculated by,

$$\mathscr{L}_R = \sum_{i,j} \theta_{ij} ||f(x_i) - f(x_j)||^2. \tag{5.2}$$

Here, θ_{ij} is a graph-based weight. The edges in the graph between each pair of data points i and j are weighted. The closer the two points are in Euclidean space d_{ij}, the greater the weight θ_{ij}. In this work, we chose a Gaussian similarity function to calculate the weights, given as follows:

$$\theta_{ij} = \exp\left(-\frac{d_{ij}^2}{\sigma^2}\right) = \exp\left(-\frac{\sum_d \left(x_i^d - x_j^d\right)^2}{\sigma^2}\right) \tag{5.3}$$

where σ is a length scale parameter. This parameter has an impact on the graph structure; hence, the value of σ needs to be selected carefully [133].

5.2.3 The Proposed Framework

We formulate the framework of fair SSL as follows, including the classification loss, the label propagation loss and fairness constraints.

$$\min_{\mathbf{w},\mathbf{y_u}} \mathscr{L}_C(\mathbf{w},\mathbf{y_u}) + \alpha \mathscr{L}_R(\mathbf{y_u}) \qquad s.t.\ \ s(\mathbf{w}) \leq c \tag{5.4}$$

where $\mathscr{L}_C$ is the classification loss between predicted labels and true labels; $\mathscr{L}_R$ is the loss of label propagation from labeled data to unlabeled data; α is a parameter to balance the loss; $s(\mathbf{w})$ is the expression of fairness constraints; and c is a threshold.

5.2.3.1 Classification Loss

A classification loss function evaluates how well a specific algorithm models the given dataset. When different algorithms are used to train datasets, such as logistic regression or neural networks, a corresponding loss function is applied to evaluate the accuracy of the model.

5.2.3.2 Label Propagation Loss

According to [132], when Laplacian regularizer is used, the label propagation loss for $\mathscr{J}_{\mathscr{L}}$ through SSL can be expressed as,

$$\mathscr{J}_{\mathscr{L}} = \min_{\mathbf{y_u}} \mathrm{Tr}(\mathbf{y}^T L \mathbf{y}) \tag{5.5}$$

where Tr denotes the trace, and the vector $\mathbf{y} = [\mathbf{y}_l; \mathbf{y}_u] \in \mathbb{R}^k$ includes labels of labeled and unlabeled data.

5.2.3.3 Fairness Constraints

Adding fairness constraints is a useful method to enforce fair learning with in-processing methods. In SSL, labeled data and unlabeled data have different impacts on discrimination because of two reasons: 1) predicting labels for unlabeled data will bring noise to the labels; 2) labeled data and unlabeled data may have different data distributions. Therefore, the discrimination inherently in unlabeled data is different from the discrimination in labeled data. For these reasons, we impose fairness constraints on labeled and unlabeled data to measure discrimination to see the disparate impact of fairness constraints on labeled and unlabeled data. We consider four cases of fairness constraints enforced on the training data:

- 1. Labeled data: $s_1(\mathbf{w}) \le c$.
- 2. Unlabeled data: $s_2(\mathbf{w}) \le c$.
- 3. Combined labeled and unlabeled data: $s_1(\mathbf{w}) \le c_1$ (for labeled data) and $s_2(\mathbf{w}) \le c_2$ (for unlabeled data).
- 4. Mixed labeled and unlabeled data: $s(\mathbf{w}) \le c$.

where c is a discrimination threshold, that is adjusted to control the trade-off between accuracy and fairness. Note that many fairness constraints [4, 145, 146] have been proposed to enforce various fairness metrics, such as disparate impact and disparate mistreatment, and these fairness constraints can be used in our framework. The basic idea to design fairness constraints is that using the covariance between the users,Äô sensitive attributes and the signed distance between the feature vectors restricts the correlation between sensitive attributes and classification results. This can be described as,

$$|\frac{1}{k}\mathbf{g_w}(\mathbf{z} - \bar{z})| \le c \tag{5.6}$$

where $\mathbf{g_w^T} \in \mathbb{R}^k$ is a vector that denotes the signed distance between the feature vectors and the decision boundary of a classifier. The form of $\mathbf{g_w}$ is different in fairness metrics, and we list them in the following,

- Disparate impact

$$\mathbf{g_w} = \mathbf{w^T}X \tag{5.7}$$

- Overall misclassification rate

$$\mathbf{g_w} = \min\left(0, \mathbf{y}^T\mathbf{y}\mathbf{w}^T X\right) \tag{5.8}$$

- False positive rate

$$\mathbf{g_w} = \min\left(0, \frac{\mathbf{1} - \mathbf{y}^T}{2}\mathbf{y}\mathbf{w}^T X\right) \tag{5.9}$$

- False negative rate

$$\mathbf{g_w} = \min\left(0, \frac{\mathbf{1}+\mathbf{y}^T}{2}\mathbf{y}\mathbf{w}^T X\right) \tag{5.10}$$

5.2.4 Fair SSL of Logistic Regression

In this section, we propose algorithms to solve the optimization problem (10) with a binary logistic regression (LR) classifier. (Other margin classifiers can also be applied in our method, and we give another example of support vector machines in the supplemental material.) The classifier is subjected to the fairness metric of demographic parity with mixed labeled and unlabeled data. The objective function of LR is defined as,

$$\mathscr{L}_C^{LR} = -\ln(\mathbf{p})\mathbf{y} - \ln(\mathbf{1}-\mathbf{p})(\mathbf{1}-\mathbf{y}) \tag{5.11}$$

where $\mathbf{p} = \frac{1}{1+e^{-\mathbf{w}^T X}}$ is the probability distribution of mapping X to the class label $\mathbf{y}$; $\mathbf{1}$ denotes a column vector with all its elements being 1. Given the logistic regression loss, the label propagation loss and the fairness metric, the optimized problem (5.5) adopts the form,

$$\begin{aligned} &\min_{\mathbf{w},\mathbf{y_u}} -\ln(\mathbf{p})\mathbf{y} - \ln(\mathbf{1}-\mathbf{p})(\mathbf{1}-\mathbf{y}) + \alpha Tr(\mathbf{y}^T L\mathbf{y}) \\ &s.t. \quad |\frac{1}{k}\mathbf{g_w}(\mathbf{z}-\bar{z})| \leq c \end{aligned} \tag{5.12}$$

5.2.4.1 Disparate impact

First, we solve the optimization problem with disparate impact as the fairness metric. The optimization of problem (5.12) includes two parts: learning the weights $\mathbf{w}$ and predicted labels of unlabeled data $\mathbf{y_u}$. The basic idea of the solution is that because of the fairness constraint, the weight $\mathbf{w}$ is updated towards a fair direction, and using the updated $\mathbf{w}$ to update $\mathbf{y_u}$ also ensures that $\mathbf{y_u}$ is directed towards fairness. The problem is solved by updating $\mathbf{w}$ and $\mathbf{y_u}$ iteratively as follows.

Solving $\mathbf{w}$ when $\mathbf{y_u}$ is fixed, the problem (5.12) becomes

$$\begin{aligned} &\min_{\mathbf{w}} -\ln(\mathbf{p})\mathbf{y} - \ln(\mathbf{1}-\mathbf{p})(\mathbf{1}-\mathbf{y}) \\ &s.t. \quad |\frac{1}{k}\mathbf{w}^T X(\mathbf{z}-\bar{z})| \leq c \end{aligned} \tag{5.13}$$

Note that problem (5.13) is a convex problem that can be written as a regularized optimization problem by moving fairness constraints to the objective function. The optimal $\mathbf{w}^*$ can then be calculated by using KKT conditions.

Solving $\mathbf{y_u}$ when $\mathbf{w}$ is fixed, the problem (5.12) becomes

$$\min_{\mathbf{y_u}} -\ln(\mathbf{p})\mathbf{y} - \ln(\mathbf{1}-\mathbf{p})(\mathbf{1}-\mathbf{y}) + \alpha Tr(\mathbf{y}^T L\mathbf{y}) \tag{5.14}$$

Given that problem (5.14) is also a convex problem, the optimal $\mathbf{y_u}$ can be obtained from the deviation of $\mathbf{y_u}$ in problem (20). In order to calculate $\mathbf{y_u}$ conveniently, we split Laplacian matrix L into four blocks after the l-th row and the l-th column: $L = \begin{bmatrix} L_{ll} & L_{lu} \\ L_{ul} & L_{uu} \end{bmatrix}$. The deviation of Eq.(5.14) is then calculated w.r.t. $\mathbf{y_u}$ and setting to zero, we have

$$\alpha(2\mathbf{y_u}L_{uu} + L_{ul}\mathbf{y_l} + (\mathbf{y_l}L_{lu})^T) - [(ln(\mathbf{p}))^T + (ln(\mathbf{1}-\mathbf{p}))^T] = 0 \tag{5.15}$$

Note that L is a symmetric matrix and, after simplification, the closed updated form of $\mathbf{y_u}$ can be derived from

$$\mathbf{y_u} = -L_{uu}^{-1}(L_{ul}\mathbf{y_l} + \frac{1}{2\alpha}[(ln(\mathbf{p}))^T + (ln(\mathbf{1}-\mathbf{p}))^T]) \tag{5.16}$$

Note that the computed optimal $\mathbf{y_u}$ is decimals, and it cannot be used to update $\mathbf{w}$ directly because only integers are allowed to optimize $\mathbf{w}$ in the next update. Due to this, we need to convert $\mathbf{y_u}$ from decimals to integers to update $\mathbf{w}$. Before using $\mathbf{y_u}$ to update the next $\mathbf{w}$, the value of $y_{u,i} \in \mathbf{y_u}, i = 1, ..., k_u$ is set to,

$$y_{u,i} = \begin{cases} 1, & y_{u,i} \geq T \\ 0, & y_{u,i} < T \end{cases} \tag{5.17}$$

where T is the threshold that determines the classification result. Then, the optimization problem (5.12) can be solved by optimizing $\mathbf{w}$ and $\mathbf{y_u}$ iteratively. **Algorithm 1** summarizes the solution of optimization problem (5.12) with the disparate impact.

5.2.4.2 Disparate mistreatment

Disparate mistreatment metrics include overall misclassification rate, false positive rate and false negative rate. For simplicity, the overall misclassification rate is used to

Algorithm 6 The algorithm of the optimizing problem (5.12) with disparate impact

```
Input: Labeled dataset $D_l$, unlabeled dataset $D_u$, fairness thresholds $c$
Parameter: $T$, $\sigma$
Initialize: Given initial values of $\mathbf{y_u}$ by label propagation
Output: $\mathbf{w}$ and $\mathbf{y_u}$
Calculate the adjacent matrix $A$ according to Eq.(5.3)
repeat
    Fix $\mathbf{y_u}$ and update $\mathbf{w}$ with KKT
    Fix $\mathbf{w}$ and update $\mathbf{y_u}$ by Eq.(5.16)
    Set $y_{u,i} \in \mathbf{y_u}$ to 0 or 1 by Eq. (5.17)
until The optimization problem (5.12) convergs
```

analyze disparate mistreatment. However, false positive rate and false negative rate can also be analyzed easily, and the result of three disparate mistreatment metrics are presented in the experiment.

With the overall misclassification rate as the fairness metric, the objective function is denoted as,

$$\begin{aligned} &\min_{\mathbf{w}} -\ln(\mathbf{p})\mathbf{y} - \ln(\mathbf{1}-\mathbf{p})(\mathbf{1}-\mathbf{y}) + \alpha Tr(\mathbf{y}^T L\mathbf{y}) \\ &s.t. \quad |\frac{1}{k}\mathbf{g_w}(\mathbf{x})(\mathbf{z}-\bar{z})| \le c \end{aligned} \tag{5.18}$$

Note that fairness constraints of disparate mistreatment are non-convex, and the solution to the optimization problem (5.18) is more challenging than the optimization problem in (5.12). Next, we convert these constraints into a Disciplined Convex-Concave Program (DCCP). Thus, the optimization problem (5.18) can be solved efficiently with the recent advances in convex-concave programming [121].

The fairness constraint of disparate mistreatment can be split into two terms,

$$\frac{1}{k}|\sum_{D_0}(0-\bar{z})\mathbf{g_w} + \sum_{D_1}(1-\bar{z})\mathbf{g_w}| \le c \tag{5.19}$$

where D_0 and D_1 are the subsets of the labeled dataset D_l and unlabeled dataset D_u with values $z = 0$ and $z = 1$, respectively. k_0 and k_1 are defined as the number of data points in the D_0 and D_1, and thus $\bar{z}$ can be rewritten as $\bar{z} = \frac{0*k_0+1*k_1}{k} = \frac{k_1}{k}$. Then the fairness constraint of disparate mistreatment can be rewritten as,

$$\frac{k_1}{k}|\sum_{D_0}\mathbf{g_w} + \sum_{D_1}\mathbf{g_w}| \le c \tag{5.20}$$

Solving $\mathbf{w}$ when $\mathbf{y_u}$ is fixed, the problem (5.18) becomes

$$\begin{aligned} &\min_{\mathbf{w}} -\ln(\mathbf{p})\mathbf{y} - \ln(\mathbf{1}-\mathbf{p})(\mathbf{1}-\mathbf{y}) \\ &s.t. \quad \frac{k_1}{k} |\sum_{D_0} \mathbf{g_w} + \sum_{D_1} \mathbf{g_w}| \leq c \end{aligned} \tag{5.21}$$

The optimization problem (5.18) is a Disciplined Convex-Concave Program (DCCP) for any convex loss, and can be solved with some efficient heuristics [121].

Solving $\mathbf{y_u}$ when $\mathbf{w}$ is fixed, the problem (5.18) becomes

$$\min_{\mathbf{y_u}} -\ln(\mathbf{p})\mathbf{y} - \ln(\mathbf{1}-\mathbf{p})(\mathbf{1}-\mathbf{y}) + \alpha Tr(\mathbf{y}^T L \mathbf{y}) \tag{5.22}$$

The solution of Eq. (5.22) is the same as the solution of Eq. (5.15). The closed form of $\mathbf{y_u}$ can be obtained via Eq. (5.16), and then the optimization problem (5.18) can be solved by updating $\mathbf{y_u}$ and $\mathbf{w}$ iteratively. **Algorithm 2** summarizes this process.

Algorithm 7 The algorithm of the optimizing problem (5.18)

```
Input: Labeled dataset $D_l$, unlabeled dataset $D_u$, fairness thresholds $c$
Parameter: $T$, $\sigma$
Initialize:Given initial values of $\mathbf{y_u}$ by label propagation
Output: $\mathbf{w}$ and $\mathbf{y_u}$
Calculate the adjacent matrix $A$ according to Eq.(5.3)
Choose a metric in disparate mistreatment
repeat
    Divide $\mathscr{D}$ into $\mathscr{D}_0$ and $\mathscr{D}_1$
    Calculate $k_0$ and $k_1$
    Fix $\mathbf{y_u}$ and update $\mathbf{w}$ with DCCP
    Fix $\mathbf{w}$ and update $\mathbf{y_u}$ by Eq.(5.16)
    Set $y_{u,i} \in \mathbf{y_u}$ to 0 or 1 by Eq. (5.17)
until The optimization problem (5.18) convergs
```

5.2.5 A Case of SVM

SVM can also be applied in our framework. SVM uses a hyperplane $\boldsymbol{w}^T X = 0$ to classify data points. The loss function of SVM is defined as,

$$\mathscr{L}_1^{SVM} = \frac{1}{K}\left(1 - \boldsymbol{y}\left(\boldsymbol{w}^T X\right)\right) \tag{5.23}$$

Based on SVM loss, the label propagation and the fairness metrics, the objective function is given as,

$$\begin{aligned} &\min_{\boldsymbol{w},\boldsymbol{y_u}} \frac{1}{K}\left(1-\boldsymbol{y}\left(\boldsymbol{w}^T X\right)\right)+\alpha Tr(\boldsymbol{y}^T L\boldsymbol{y}) \\ &s.t. \quad |\frac{1}{K}g_w(\boldsymbol{z}-\bar{z})| \leq c \end{aligned} \tag{5.24}$$

Disparate impact and disparate mistreatment can be used in the fairness constraint. Since SVM loss is convex, the solution to the problem (5.24) is similar to the LR case. For simplicity, we omit the process and show the results of the experiment.

5.2.6 Discussion

Based on the above analysis, some conclusions can be drawn:

1. Since unlabeled data do not contain any label information, they do not label biased information so that we can take advantage of the unlabeled data to improve the trade-off between accuracy and fairness. In our framework, due to the fairness constraint, the weight **w** is updated towards a fair direction. Using the updated **w** to update $\mathbf{y_u}$ also ensures that $\mathbf{y_u}$ is directed towards fairness. In this way, fairness is enforced in labeled and unlabeled data by updating **w** and $\mathbf{y_u}$ iteratively. Therefore, labels of unlabeled data are calculated in a fair way, which is beneficial to the accuracy of the classifier as well as the fairness of the classifier.

2. Fairness constraints on labeled data and unlabeled data have a different impact on the training result because labeled and unlabeled data may present different covariance between the sensitive attribute and the signed distance between feature vectors to the decision boundaries.

5.3 Experiment

In this section, we first describe the experimental setup, including datasets, baselines, and parameters. Then, we evaluate our method on three real-world datasets under the fairness metric of disparate impact and disparate mistreatment (including OMR, FNR and FPR). The aim of our experiments is to assess: the effectiveness of our method to achieve fair semi-supervised learning; the impact of different fairness constraints on fairness; and the extent to which unlabeled can balance fairness with accuracy.

5.3.1 Experimental Setup

5.3.1.1 Dataset

Our experiments involve three real-world datasets: Health dataset [1], Titanic dataset [2] and Bank dataset [3].

- The task in the Health dataset is to predict whether people will spend time in the hospital. In order to convert the problem into the binary classification task, we only predict whether people will spend any day in the hospital. After data preprocessing, the dataset contains 27,000 data points with 132 features. We divide patients into two groups based on age (≥65 years) and consider 'Age' to be the sensitive attribute.

- The Bank dataset contains a total of 41,188 records with 20 attributes and a binary label, which indicates whether the client has subscribed (positive class) or not (negative class) to a term deposit. We consider 'Age' as sensitive attribute.

- The Titanic dataset comes from a Kaggle competition where the goal is to analyze which sorts of people were likely to survive the sinking of the Titanic. We consider "Gender" as the sensitive attribute. After data preprocessing, we extract 891 data points with 9 features.

5.3.1.2 Parameters

The sensitive attributes are excluded from the training set to ensure fairness between groups and are only used to evaluate discrimination in the test phrases. In the Health, Bank and Titanic datasets, data are all labeled. In the Health dataset, we sample 4000 data points as labeled dataset, 4000 data points as test dataset, and left as unlabeled dataset. In the Bank dataset, we sample 4000 data points as labeled dataset, 4000 data points as test dataset, and left as unlabeled dataset. In the Titanic dataset, we sample 200 data points as labeled dataset, 200 data points as test dataset, and left as unlabeled dataset. Therefore, $\mathscr{D}_l$ and $\mathscr{D}_u$ are collected from the similar data distribution.

[1] https://foreverdata.org/1015/index.html

[2] https://www.kaggle.com/c/titanic/data

[3] https://archive.ics.uci.edu/ml/datasets/bank+marketing

In the experiments, the results are an average of 10 results by randomly sampling labeled dataset, test dataset and unlabeled dataset. We set $\alpha = 1$ and $T = 0.5$ in all datasets; $\sigma = 0.5$ in the Health dataset and Bank dataset, and $\sigma = 0.1$ in the Titanic dataset. In DCCP, parameter τ is set to 0.05 and 1 in the Bank and Titanic dataset, respectively. Parameter μ is set to 1.2 in both Bank and Titanic datasets.

5.3.1.3 Baseline Methods

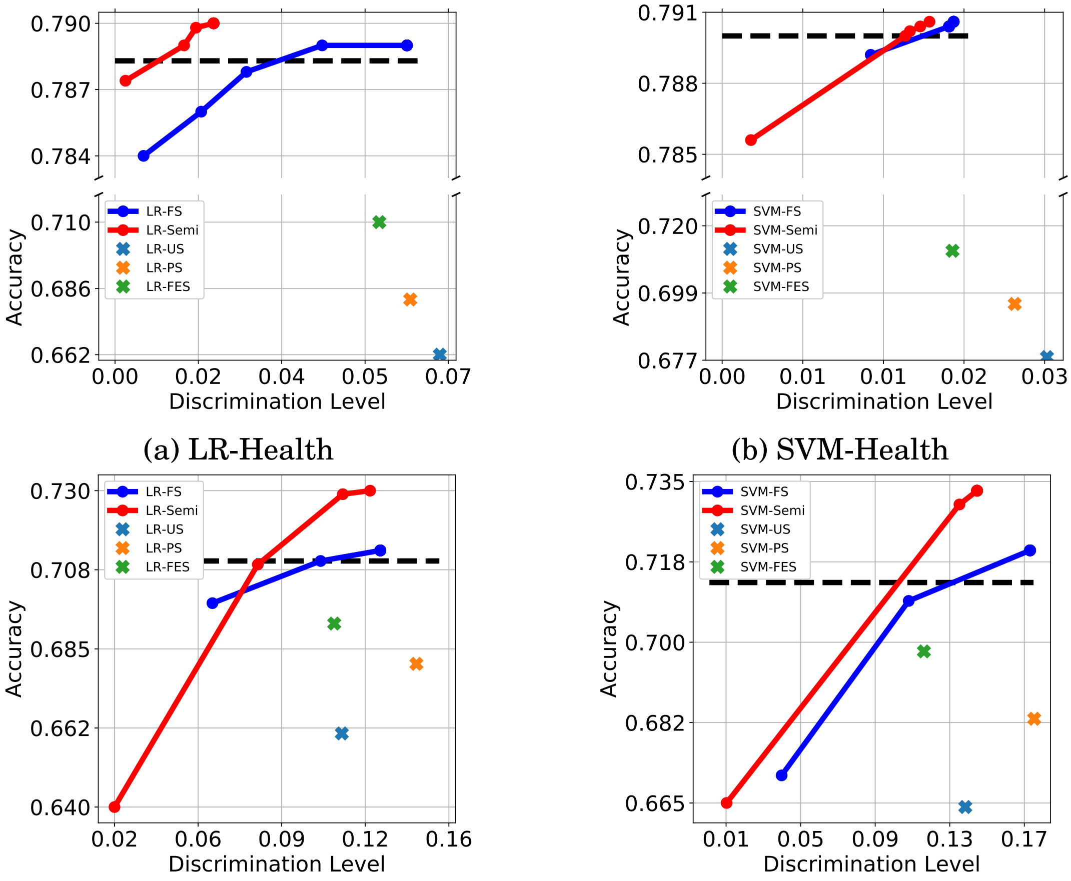
The methods chosen for comparison are listed as follows. It is worth noting that [32] also used unlabeled data on fairness. However, they only applied the equal opportunity metric, which is different to ours. Hence, we did not compare the proposed method with them.

- Fairness Constraints (FS): Fairness constraints are used to ensure fairness for classifiers. [146]
- Uniform Sampling (US): The number of data points in each groups is equalized through oversampling and/ undersampling. [73]
- Preferential Sampling (PS): The number of data points in each groups is equalized by taking samples near the borderline data points. [73]
- Fairness-enhanced sampling (FES): A fair SSL framework includes pseudo labeling, re-sampling and ensemble learning. [149]

5.3.2 Experimental Results of Disparate Impact

5.3.2.1 Trade-off Between Accuracy and Discrimination

Figure 5.1 shows that as c varies, accuracy and discrimination level in the proposed method and other methods with LR and SVM on two datasets. From the results, we can observe that our framework provides a better trade-off between accuracy and discrimination. A better trade-off means that with the same accuracy, discrimination is low or with the same discrimination, accuracy is higher. For example, at the same level of accuracy on the Titanic dataset, (shown by the black line), our method with LR has a discrimination level of around 0.08, while the FS method has a discrimination



(a) LR-Health

(b) SVM-Health

(c) LR-Titanic

(d) SVM-Titanic

Figure 5.1: The trade-off between accuracy and discrimination in proposed method Semi (Red), FS (Blue), US (Blue cross), PS (Yellow cross) and FES (Green cross) under the fairness metric of disparate impact with LR and SVM in two datasets. As the threshold of covariance c increases, accuracy and discrimination increase. The results demonstrate that our method achieves a better trade-off between accuracy and discrimination than other methods.

level of 0.11. A similar observation can be made from the results with the PS method (Yellow cross), US method (Blue cross) and the FES method (Green cross). Note that the discrimination level (red line) with LR in the Health dataset does not extend because discrimination does not increase as c grows. Additionally, we note that accuracy and discrimination level are related to the training models. In the Titanic dataset, LR has lower accuracy and discrimination than SVM and the choice of training models is related to the datasets.

Dataset	Health dataset							
Constraint	Labeled		Unlabeled		Combined		Mixed	
	Acc	Dis	Acc	Dis	Acc	Dis	Acc	Dis
c=0.0	0.7868	0.0042	N/A	N/A	0.7874	0.0022	0.7862	0.0003
c=0.1	0.7890	0.0129	N/A	N/A	0.7890	0.0145	0.7892	0.0149
c=0.2	0.7900	0.0170	0.7900	0.0207	0.7898	0.0170	0.7898	0.0170
c=0.3	0.7898	0.0207	0.7898	0.0170	0.7900	0.0207	0.7900	0.0207
c=0.4	0.7902	0.0178	0.7898	0.0170	0.7900	0.0207	0.7900	0.0207
c=0.5	0.7900	0.0207	0.7900	0.0207	0.7900	0.0207	0.7900	0.0207
c=0.6	0.7900	0.0207	0.7906	0.0186	0.7900	0.0207	0.7900	0.0207
c=0.7	0.7900	0.0207	0.7900	0.0207	0.7900	0.0207	0.7900	0.0207
c=0.8	0.7900	0.0207	0.7904	0.0191	0.7900	0.0207	0.7900	0.0207
c=0.9	0.7900	0.0207	0.7908	0.0190	0.7900	0.0207	0.7900	0.0207
c=1.0	0.7900	0.0207	0.7900	0.0207	0.7900	0.0207	0.7900	0.0207

Table 5.1: The impact of fairness constraints on different datasets in terms of accuracy (Acc) and discrimination level (Dis) under the fairness metric of disparate impact with LR in the Health dataset.

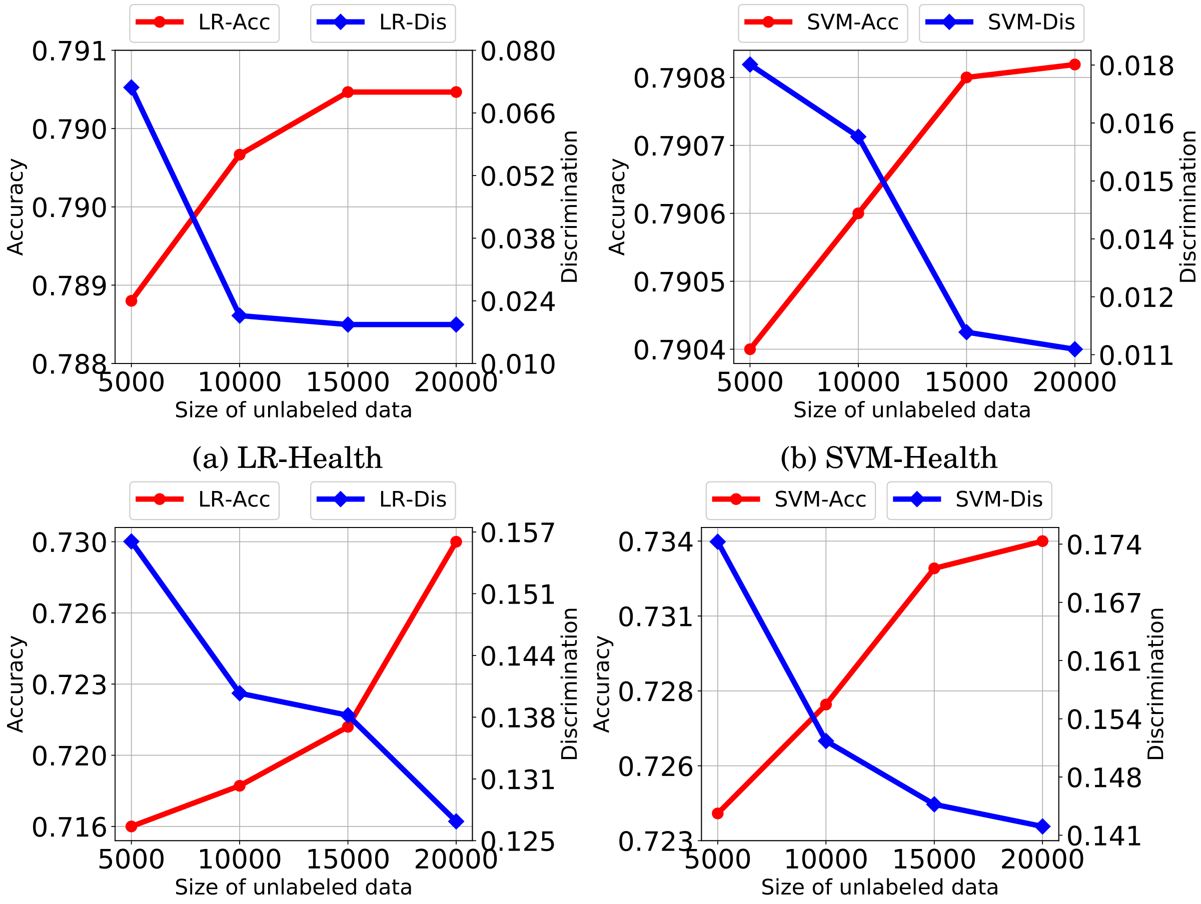
Dataset	Titanic dataset							
Constraint	Labeled		Unlabeled		Combined		Mixed	
	Acc	Dis	Acc	Dis	Acc	Dis	Acc	Dis
c=0.0	0.6330	0.0128	0.6970	0.1244	0.6290	0.0139	0.6440	0.0402
c=0.05	0.6690	0.0579	0.7070	0.1265	0.6690	0.0716	0.6810	0.0948
c=0.1	0.7150	0.1272	0.7140	0.1332	0.7100	0.1239	0.7150	0.1256
c=0.15	0.7200	0.1366	0.7190	0.1336	0.7190	0.1336	0.7200	0.1366
c=0.2	0.7200	0.1366	0.7200	0.1366	0.7200	0.1366	0.7200	0.1366
c=0.25	0.7200	0.1366	0.7200	0.1366	0.7200	0.1366	0.7200	0.1366

Table 5.2: The impact of fairness constraints on different datasets in terms of accuracy (Acc) and discrimination level (Dis) under the fairness metric of disparate impact with LR in the Titanic dataset.

5.3.2.2 Different Fairness Constraints

Our next set of experiments is to determine the impact of different fairness constraints. For these tests, the size of unlabeled data is set to 12,000 data points in the Health dataset and 400 data points in the Titanic dataset. Due to space limitations, we have only reported the results for the LR, which appear in Tables 5.1 and 5.2. The result shows that when varying the threshold of covariance c, different fairness constraints on labeled and unlabeled data have different impacts on the training results. As the threshold of covariance increases, both accuracy and discrimination level increase before steadying off for the duration. In terms of accuracy, this is because a larger c allows for a larger space to find better weights $\boldsymbol{w}$ to inform classification. In terms of discrimination, a larger c tends to introduce more discrimination in noise.

It is also observed that the fairness constraint on mixed data generally has the best performance in the trade-off between accuracy and discrimination. The other three constraints have very similar accuracy and discrimination levels. We attribute this to the assumption that labeled and unlabeled data have a similar data distribution, and therefore the mixed fairness constraint on labeled and unlabeled data gives the best description of the covariance between sensitive attributes and signed distance from feature vectors to the decision boundary.



(a) LR-Health

(b) SVM-Health

(c) LR-Titanic

(d) SVM-Titanic

Figure 5.2: The impact of the amount of unlabeled data in the training set on accuracy (Red) and discrimination level (Blue) under the fairness metric of disparate impact with LR and SVM in two datasets. The X-axis is the size of unlabeled dataset; the left y-axis is accuracy, and right y-axis is discrimination level.

5.3.2.3 The Impact of Unlabeled Data

For these experiments, we set the covariance threshold $c = 1$ for the Health and Titanic datasets. Figure 5.2 shows that accuracy and discrimination level vary with the amount of unlabeled data. This applies to both the LR and SVM classifiers on both datasets. As shown, accuracy increases as the amount of unlabeled data increases in both datasets before stabilizing at its peak. Discrimination level sharply decreases almost immediately, then also stabilize. These results clearly demonstrate that discrimination in variance decreases as the amount of unlabeled data in the training set increases.

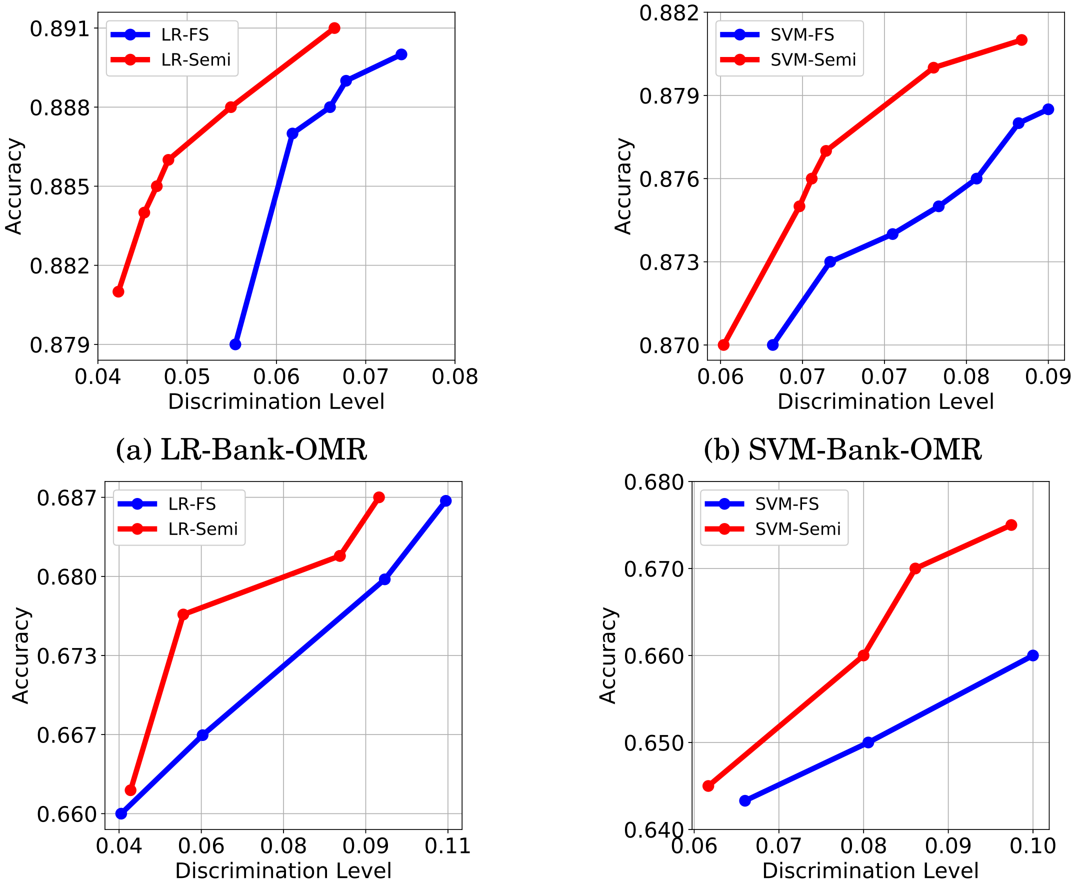
5.3.3 Experimental Results of Disparate Mistreatment

5.3.3.1 Trade-off Between Accuracy and Discrimination

Figures 5.3-5.5 show that as c varies, accuracy and discrimination level in the proposed framework and the FS method with LR and SVM on two datasets under the fairness metric of OMR, FPR and FNR. From the results, we can observe that our proposed method (Red line) generally is on the left above the FS method (Blue line). This indicates that our framework provides a better trade-off between accuracy and discrimination in three metrics of most time. For example, at the same level of accuracy (Acc = 0.885) on the Bank dataset under OMR, our method with LR has a discrimination level of around 0.045, while the FS method has a discrimination level of 0.06. We also observe that discrimination level is quite different under fairness metrics. For example, discrimination level can reach 0.17 at the end under FNR, while discrimination level only shows 0.01 under FPR. In addition, we note that accuracy and discrimination level have different performances on training models. In the Bank dataset, SVM generally has lower accuracy and discrimination than LR.

5.3.3.2 Different Fairness Constraints under OMR

Table 5.3 and Table 5.4 shows that different fairness constraints on labeled and unlabeled data have different impacts on the training results. Due to space limitations, we have only reported the results for the LR under the metric of OMR on the Bank and Titanic datasets. For these tests, the size of unlabeled data is set to 4,000 data points in the Bank dataset and 400 data points in the Titanic dataset. As shown, when varying



(a) LR-Bank-OMR

(b) SVM-Bank-OMR

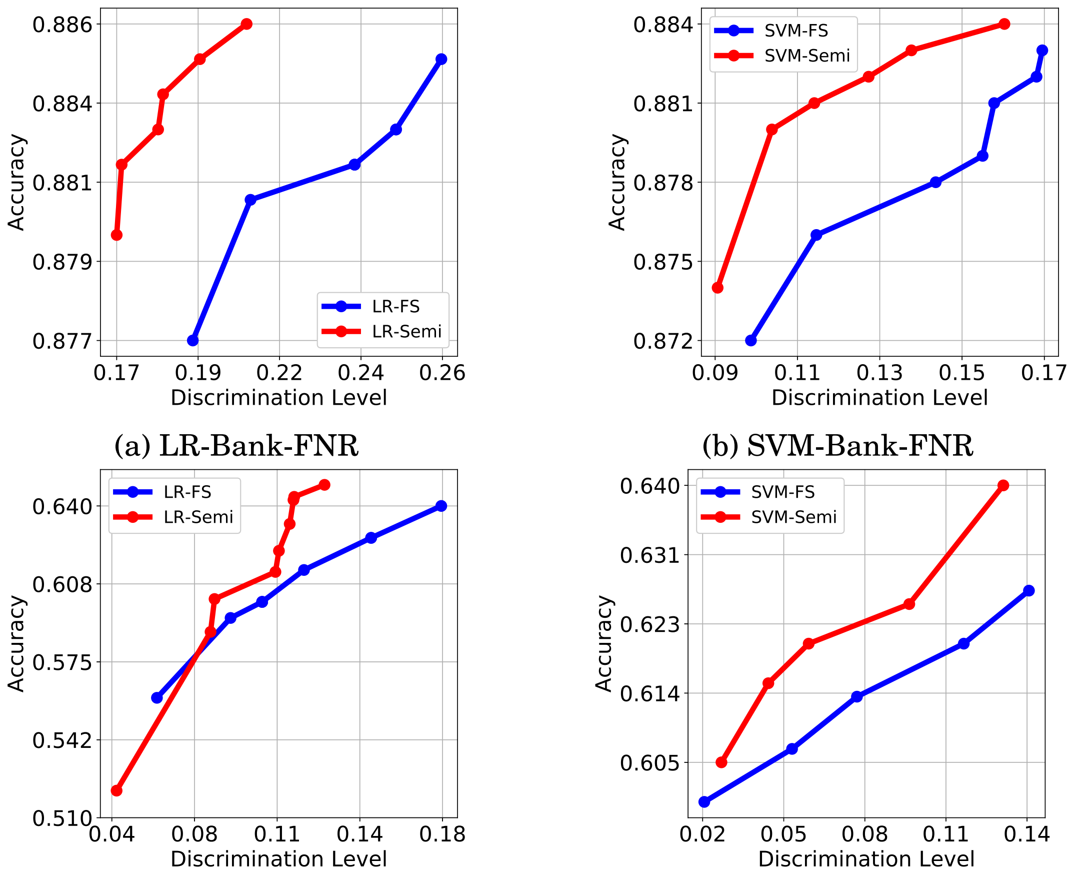
(c) LR-Titanic-OMR

(d) SVM-Titanic-OMR

Figure 5.3: The trade-off between accuracy and discrimination in proposed method Semi (Red), FS (Blue) with LR and SVM in two datasets under the metric of overall misclassification rate. As the threshold of covariance c increases, accuracy and discrimination increase. The results demonstrate that our method using unlabeled data achieves a better trade-off between accuracy and discrimination.

the threshold of covariance c, different fairness constraints on labeled and unlabeled data have a huge difference in the training results. When the fairness constraint is enforced in labeled data, accuracy and discrimination increase with the increase in c in the Titanic dataset. This is because a smaller c enforces the lowest discrimination level, which results in lower accuracy.

However, when the fairness constraint is enforced in unlabeled data, accuracy and discrimination could decrease with the increase in c. This is because the label of



(a) LR-Bank-FNR

(b) SVM-Bank-FNR

(c) LR-Titanic-FNR

(d) SVM-Titanic-FNR

Figure 5.4: The trade-off between accuracy and discrimination in the proposed method Semi (Red), FS (Blue) with LR and SVM in two datasets under the metric of false negative rate. As the threshold of covariance c increases, accuracy and discrimination increase.

unlabeled data appears in the fairness constraint of disparate mistreatment, and it is updated during the training. This means that the distribution of unlabeled data is not described well during the training. As a result, the fairness constraint on unlabeled data is not that effective.

5.3.3.3 The Impact of Unlabeled Data under OMR

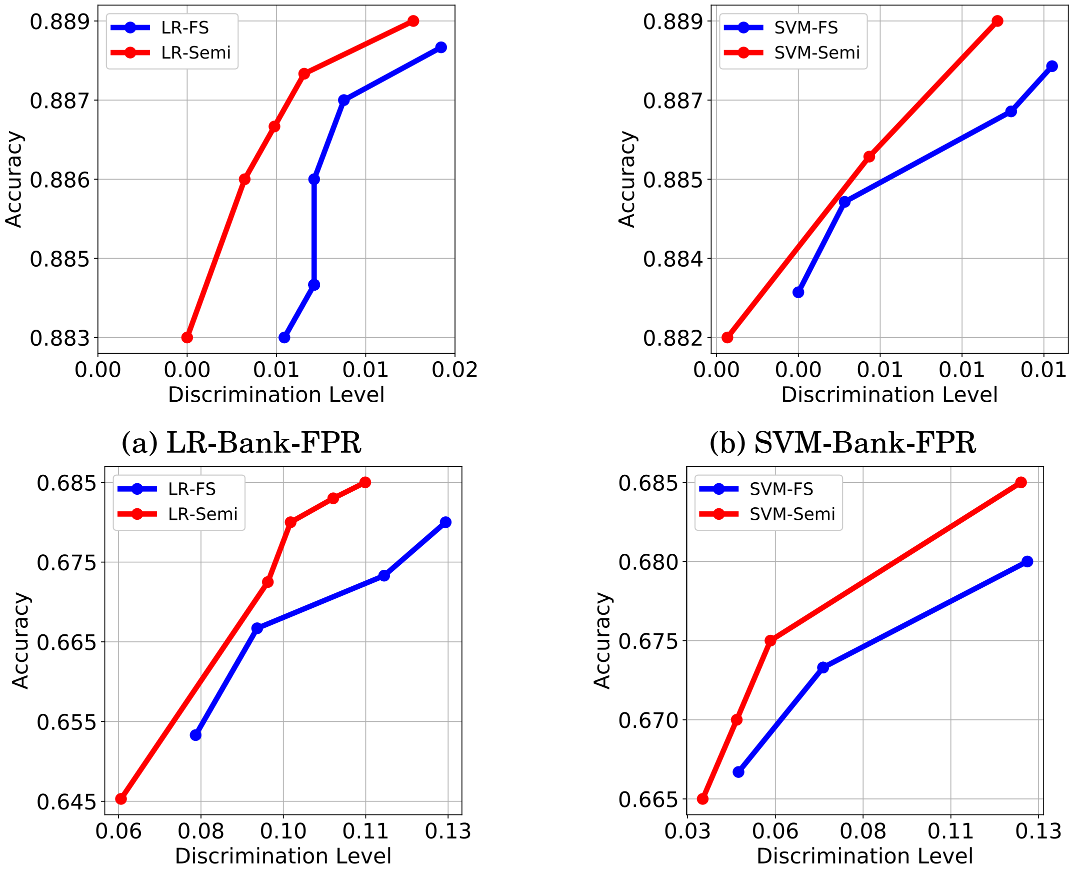
For these experiments, we show the impact of unlabeled data on OMR. The covariance threshold is set as $c = 1$ for the Bank and Titanic datasets. Figure 5.6 shows accuracy

Dataset				Bank dataset				
Constraint	Labeled		Unlabeled		Combined		Mixed	
	Acc	Dis	Acc	Dis	Acc	Dis	Acc	Dis
c=0.0	0.8635	0.0905	0.8407	0.1847	0.8342	0.147	0.8605	0.0987
c=0.5	0.8625	0.092	0.8402	0.1854	0.8342	0.1442	0.8605	0.0987
c=1.0	0.8638	0.0922	0.8402	0.1854	0.835	0.1452	0.8635	0.1071
c=1.5	0.8645	0.0918	0.8407	0.1833	0.835	0.1452	0.8635	0.1071
c=2.0	0.8648	0.0907	0.841	0.1822	0.8347	0.1462	0.8625	0.1071
c=2.5	0.8652	0.0914	0.8413	0.1812	0.8353	0.1469	0.8635	0.1084
c=3.0	0.866	0.0923	0.8413	0.1784	0.8342	0.147	0.8627	0.1084
c=3.5	0.8662	0.0927	0.8407	0.1805	0.8342	0.147	0.8627	0.1097
c=4.0	0.8665	0.093	0.841	0.1795	0.8342	0.147	0.8627	0.1097
c=4.5	0.8668	0.0919	0.8407	0.1791	0.835	0.1452	0.8635	0.1113
c=5.0	0.867	0.0909	0.8407	0.1791	0.8355	0.1444	0.8635	0.1113

Table 5.3: The impact of fairness constraints on different datasets in terms of accuracy (Acc) and discrimination level (Dis) under the fairness metric of overall misclassification rate with LR in the Bank dataset.

Dataset				Titanic dataset				
Constraint	Labeled		Unlabeled		Combined		Mixed	
	Acc	Dis	Acc	Dis	Acc	Dis	Acc	Dis
c=0.0	0.7448	0.0285	0.7138	0.3996	0.7655	0.1387	0.7483	0.0175
c=0.5	0.7483	0.0335	0.6966	0.4386	0.7655	0.1547	0.7483	0.0175
c=1.0	0.7517	0.0385	0.6931	0.4656	0.7552	0.1397	0.7517	0.0225
c=1.5	0.7552	0.0436	0.7103	0.3946	0.7793	0.1748	0.7448	0.0445
c=2.0	0.7552	0.0436	0.7069	0.4216	0.7724	0.1648	0.7483	0.0495
c=2.5	0.7586	0.0326	0.7103	0.4106	0.7759	0.1378	0.7448	0.0605
c=3.0	0.7552	0.0596	0.7552	0.2678	0.7552	0.0596	0.7483	0.0655
c=3.5	0.7552	0.0596	0.6931	0.4656	0.7552	0.0596	0.7483	0.0816
c=4.0	0.7586	0.0646	0.7103	0.4106	0.7586	0.0646	0.7517	0.0866
c=4.5	0.7586	0.0646	0.7138	0.3996	0.7586	0.0646	0.7517	0.0866
c=5.0	0.7552	0.0756	0.7103	0.4106	0.7552	0.0756	0.7483	0.0816

Table 5.4: The impact of fairness constraints on different datasets in terms of accuracy (Acc) and discrimination level (Dis) under the fairness metric of overall misclassification rate with LR in the Titanic dataset.



(a) LR-Bank-FPR

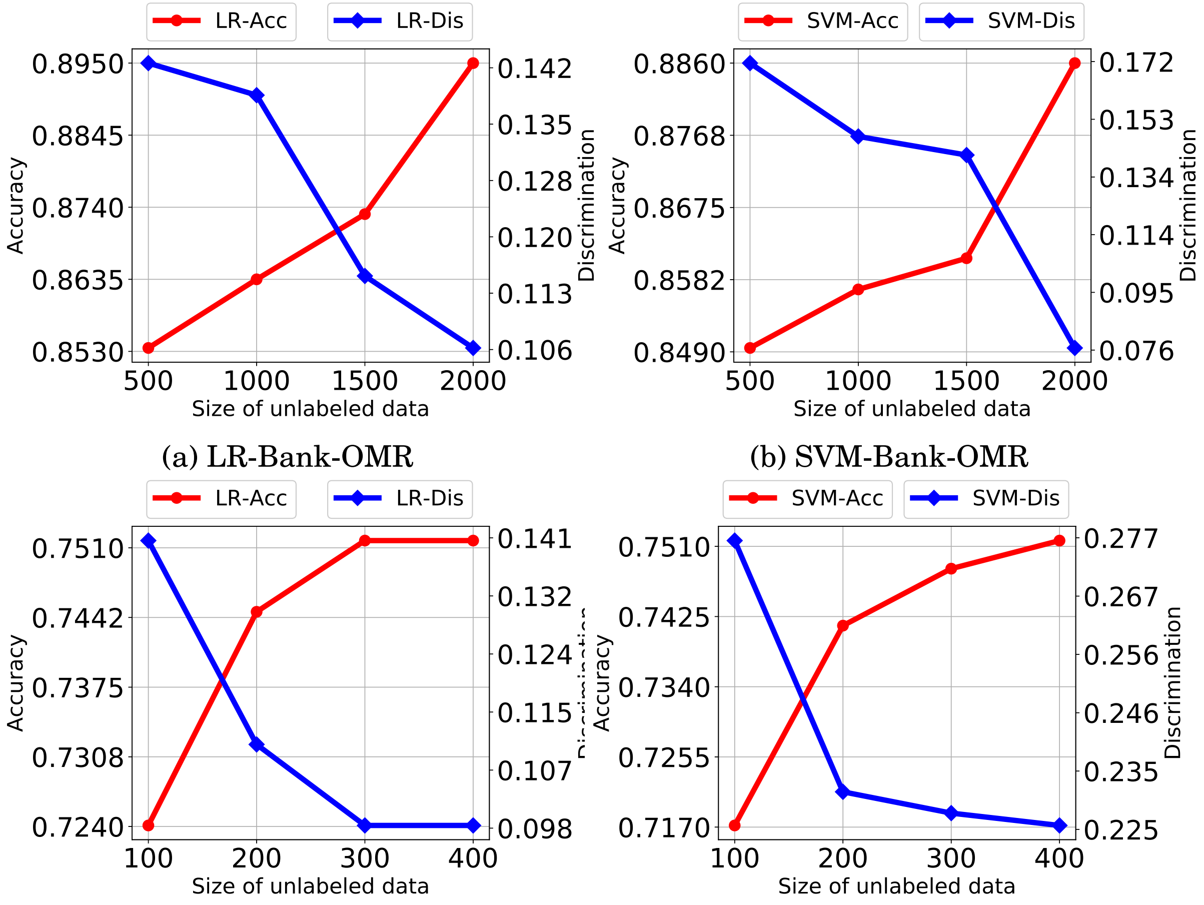
(b) SVM-Bank-FPR

(c) LR-Titanic-FPR

(d) SVM-Titanic-FPR

Figure 5.5: The trade-off between accuracy and discrimination in proposed method Semi (Red), FS (Blue) with LR and SVM in two datasets under the metric of false positive rate. As the threshold of covariance c increases, accuracy and discrimination increase.

and discrimination level varies given the different size of unlabeled data with LR and SVM on two datasets. As shown, before the peak is reached, as the amount of unlabeled data increases in the two data sets, accuracy will also increase. Discrimination level decreases at the beginning, and then stabilizes in the Titanic dataset. These results indicate that discrimination in variance decreases as the amount of unlabeled data in the training set increases.



(a) LR-Bank-OMR

(b) SVM-Bank-OMR

(c) LR-Titanic-OMR

(d) SVM-Titanic-OMR

Figure 5.6: The impact of the amount of unlabeled data in the training set on accuracy (Red) and discrimination level (Blue) under the fairness metric of overall mistreatment rate with LR and SVM in two datasets. The X-axis is the size of unlabeled dataset; the left y-axis is accuracy, and the right y-axis is discrimination level.

5.3.4 Discussion and Summary

5.3.4.1 Discussion

We discuss how the proposed framework is able to reduce discrimination in terms of discrimination decomposition into discrimination in bias, variance and noise. Discrimination in bias depends on the model choice. Discrimination in variance relates to the size of training data. Our framework uses unlabeled data to expand the size of training data, and thus reduce the discrimination in variance. Discrimination in noise depends on the quality of training data. In our framework, discrimination in noise also depends on label propagation. Predicting labels for unlabeled data may bring discrimination in noise. This discrimination can be adjusted by the threshold c in the fairness constraint. A smaller threshold generates smaller discrimination in noise. The reduction in discrimination in variance is generally more than the discrimination in noise induced by label propagation. Thus, when the same model is used, unlabeled data helps to lessen discrimination.

5.3.4.2 Summary

From these experiments, we can obtain some conclusions. 1) The proposed framework can make use of unlabeled data to achieve a better trade-off between accuracy and discrimination. 2) Under the fairness metric of disparate impact, the fairness constraint on mixed labeled and unlabeled data generally has the best trade-off between accuracy and discrimination. Under the fairness metric of disparate mistreatment, the fairness constraint on labeled data is used to achieve the trade-off between accuracy and discrimination. 3) More unlabeled data generally helps to make a better compromise between accuracy and discrimination. 4) Model choice can affect the trade-off between accuracy and discrimination. Our experiments show that SVM is more friendly to achieving a better trade-off than LR.

5.4 Related Work

In recent years, a large number of works have studied fairness in machine learning, and we classify them into two streams.

5.4.1 Fair Supervised Learning

Methods for fair supervised learning include pre-processing, in-processing and post-processing methods. In pre-processing, discrimination is eliminated by guiding the distribution of training data towards a fairer direction [73] or by transforming the training data into a new space [16, 52, 124, 147, 151]. The main advantage of the pre-processing method is that it does not require changes to the machine learning algorithm, so it is very simple to use. In in-processing, discrimination is constrained by fair constraints or a regularizer during the training phase. For example, Zafar et al. [75] used regularizer term to penalize discrimination in the learning objective. [57, 145, 146] designed the convex fairness constraint, called decision boundary covariance to achieve fair classification for classifiers. Recent work presented the constrained optimization problem as a two-player game and formalized the definition of fairness as a linear inequality [4, 36, 43]. This category is more flexible for optimizing different fair constraints, and solutions using this method are considered to be the most robust. In addition, these methods have shown good results in terms of accuracy and fairness. A third approach to achieving fairness is post-processing, where a learned classifier is modified to adjust the decisions to be non-discriminatory for different groups [60, 79, 91]. Post-processing does not need changes in the classifier, but it cannot guarantee an optimal classifier.

5.4.2 Fair Unsupervised Learning

Chierichetti et al. [30] were the first to study fairness in clustering problems. Their solution, under both k-center and k-median objectives, required every group to be (approximately) equally represented in each cluster. Many subsequent works have since been undertaken on the subject of fair clustering. Among these, Rosner et al. [116] extended fair clustering to more than two groups. Chen et al. [118] consider the fair k-means problem in the streaming model, define fair coresets and show how to compute them in a streaming setting, resulting in a significant reduction in the input size. Bera et al. [11] presented a more generalized approach to fair clustering, providing a tunable notion of fairness in clustering. Kleindessner et al. [80] studied a version of constrained spectral clustering incorporating the fairness constraints.

5.4.3 Comparing with Other Work

Existing fair learning methods mainly focus on supervised and unsupervised learning, and cannot be directly applied to SSL. As far as we know, only [32, 104, 149] has explored fairness in SSL. Chzhen et al. [32] studied Bayes classifier under the fairness metric of equal opportunity, where labeled data is used to learn the output conditional probability, and unlabeled data is used to calibrate the threshold in the post-processing phase. However, unlabeled data is not fully used to eliminate discrimination, and the proposed method only applies to equal opportunity. In [104], the proposed method is built on neural networks for SSL in the in-processing phase, where unlabeled data is marked labels with pseudo labeling. Zhang et al. [149] proposed a pre-processing framework which includes pseudo labeling, re-sampling and ensemble learning to remove representation discrimination. Our solution will focus on margin-based classifier in the in-processing stage, as in-processing methods have demonstrated good flexibility in both balancing fairness and supporting multiple classifiers and fairness metrics.

5.5 Summary

In this chapter, we propose a framework of fair semi-supervised learning that operates during in-processing phase. Our framework is formulated as an optimization problem with the goal of finding weights and labeling unlabeled data by minimizing the loss function subject to fairness constraints. We analyze several different cases of fairness constraints for their effects on the optimization problem plus the accuracy and discrimination level in the results. Our experiments confirm this analysis, showing that the proposed framework provides accuracy and fairness at high levels in semi-supervised settings.

Chapter 6

Balancing Learning Model Privacy, Fairness, and Accuracy with Early Stopping Criteria

In previous chapters, we have studied privacy and fairness in machine learning separately. In some cases, privacy, fairness, and accuracy need to be considered together when using machine learning models. For example, in disease diagnosis, demographic groups (such as black and white defendants) should experience similar processing, i.e., similar predictive accuracy. Meanwhile, participating in training data means that the individual might have been diagnosed with a disease, which is very sensitive and needs to be kept confidential. Therefore, it is crucial to consider privacy, fairness and accuracy simultaneously. In this chapter, we focus on privacy and fairness and investigate their interactions. We propose two early stop criteria to guide when to stop training to achieve a better balance between accuracy, fairness and privacy.

6.1 Introduction

Striking a balance between accuracy and privacy or accuracy and fairness are well-studied topics. Arguably, the current state-of-the-art in private machine learning is differential privacy (DP) [25, 56, 144]. With deep learning models, one of the most widely-used approaches to implementing DP is to clip gradients and add noise during

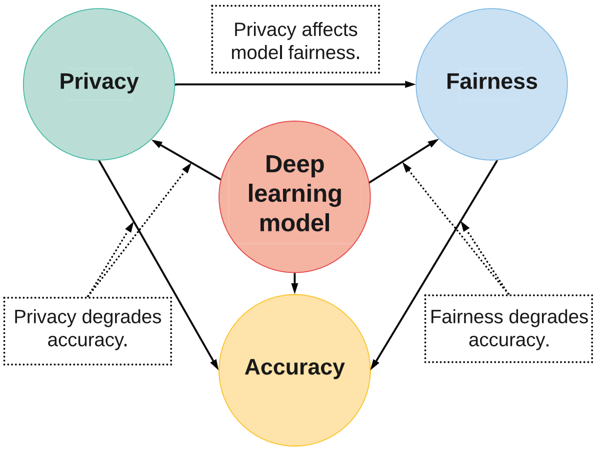

Privacy affects model fairness.
Privacy
Fairness
Deep learning model
Privacy degrades accuracy.
Fairness degrades accuracy.
Accuracy

Figure 6.1: Challenges in private and fair deep learning: 1) Privacy comes at the cost of accuracy; 2) Fairness comes at the cost of accuracy; 3) Privacy could affect model fairness.

training with DP-SGD [1]. In terms of ensuring fair machine learning, there are three broad approaches based more on when the method is applied. One type is applied before training [16, 73], the second during training [4, 145], and the third after training [60, 91]. A few studies have combined DP with fairness methods, such as the in-processing methods outlined in [40, 139], and the post-processing method outlined in [67]. However, all these techniques, whether combined with privacy or not, were designed for traditional machine learning environments. To the best of our knowledge, this is the first study to attempt a solution for balancing the trinity of fairness, privacy, and accuracy in deep learning.

In this new sphere, three big challenges await. As illustrated in Figure 6.1, all three hinge on the fact that both privacy and fairness require a sacrifice in accuracy. However, an interesting observation of DP-SGD is that DP may have a disparate impact on model accuracy. Studies by [9, 138] have shown that the reduction in accuracy caused by DP-SGD disproportionately affects under-represented groups. In other words, DP-SGD could affect model fairness. This motivated us to think that perhaps fairness does not need to be addressed directly. Rather, fairness might be better managed by treating it as a byproduct of training with DP-SGD.

In this chapter, we propose two early stop criteria to guide when to stop training, in order to achieve the ideal balance between the three. This idea comes from early stopping, which is a simple and effective method to stop training when the validation accuracy starts to decline. When using DP-SGD, gradient clipping and noise addition in DP-SGD make the training less stable than SGD. This change provides the possibility that the model may be updated to perform low discrimination levels in certain epochs. With theoretical and empirical analysis, we have noticed that the discrimination level shows similar changes on the validation set and test set. Therefore, the discrimination level in the validation set is used as a key indicator to guide when to stop training. As for privacy, when DP-SGD is used to training neural networks, less training epochs consume less privacy cost. That is to say, early stopping naturally is beneficial to save privacy cost, thereby increasing the privacy level. Thus, we conjecture that an effective way to balance accuracy, privacy and fairness is to simply stop the training process once the desired ratio between all three has been reached.

In summary, the contributions of this chapter are listed below.

- 1) We proposed two stopping criteria to guide when to stop the training. These stopping criteria help decrease the discrimination level and privacy cost without much sacrifice in accuracy.
- 2) We theoretically analyze the effect of gradient clipping and noise addition in DP-SGD on the gradient, the effect on training stability, and the effect on model accuracy and model fairness, and explain why DP-SGD affects model fairness.
- 3) We empirically analyze the impact of DP-SGD on model accuracy and fairness with both structured (Bank and Health datasets) and unstructured datasets (UTK Face and IMDB datasets). The results verify that the proposed stopping criteria help to achieve an ideal balance between accuracy, fairness and privacy.

6.2 Background

6.2.1 Notation

Consider a binary classifier $f: X \rightarrow Y$, mapping training examples x to a discrete label $y \in \{0,1\}$. Let D be a dataset with N data points $x_1, x_2, ..., x_N$, where each data point x_i contains the information of a number of d unprotected attributes $a_1, a_2, ..., a_d$, the protected attribute z, and the label y. The dataset can be divided into different groups according to the protected attribute. For example, if the protected attribute is 'Sex', the dataset can be divided into male and female groups. The training goal is to learn a mapping f_w over the dataset D to well approximately mapping between input X and output Y in a private and fair way, where w denotes the model parameters. A loss function $\mathscr{L}(f_w(x), y)$ is used to measure the difference between the label y and the classifier's prediction $f_w(x)$. The optimized model parameters w are solved by minimizing empirical loss $\mathscr{L}_D(f_w(x), y) = \frac{1}{N}\sum_{i=1}^{N} \ell_i(f_w(x), y)$. In each iteration, l, a batch of data points b is sampled from D and the gradient of the average loss ∇w^l is updated by the following rule

$$w^{l+1} = w^l - r\nabla w^l, \tag{6.1}$$

where r is the learning rate. Given a number of T training epochs, the number of iterations is $L = \frac{TN}{b}$.

6.2.2 Fairness Metrics

Definition 6.1. (Demographic parity) [15] A classification model satisfies demographic parity if

$$Pr(\hat{y} = 1|z = 1) = Pr(\hat{y} = 1|z = 0), \tag{6.2}$$

where $\hat{y}$ is the predicted label.

Definition 6.2. (Equal opportunity) [60] Equal opportunity requires that the true positive rates are equal in different groups, which is presented as

$$Pr(\hat{y} = 1|z = 1, y = 1) = Pr(\hat{y} = 1|z = 0, y = 1). \tag{6.3}$$

Definition 6.3. (Equal odds) [60] Equal odds requires that the predicted result of a classifier is independent of the sensitive attribute on the condition of true positive rate and false positive rate, which is given as

$$Pr(\hat{y}=1|z=1,y=0)=Pr(\hat{y}=1|z=0,y=0), \tag{6.4}$$

$$Pr(\hat{y}=1|z=1,y=1)=Pr(\hat{y}=1|z=0,y=1). \tag{6.5}$$

Definition 6.4. (Discrimination level) Let $\gamma_z(D,f_w)$ denote the probability of positive predictions of group z on a model f_w training with a dataset D in terms of a fairness metric. The discrimination level $\Gamma(D,f_w)$ on a model f_w training with a dataset D is measured by the difference between groups:

$$\Gamma(D,f)=|\gamma_0(D,f_w)-\gamma_1(D,f_w)|. \tag{6.6}$$

Take demographic parity as an example, we have $\gamma_1(D,f_w)=Pr(\hat{y}=1|z=1)$, and the discrimination level is $\Gamma(D,f_w)=|Pr(\hat{y}=1|z=1)-Pr(\hat{y}=1|z=0)|$.

6.2.3 Differential Privacy

DP is a widely used privacy model in machine learning to bound the leakage about the presence of specific data point [25].

Definition 6.5. (ϵ,δ-Differential privacy) [44] Given neighboring datasets D and D' that differ in one data point, an algorithm $\mathscr{M}$ satisfies (ϵ,δ)-differential privacy for any possible outcome Ω,

$$Pr[\mathscr{M}(D)\in\Omega]\le\exp(\epsilon)\cdot Pr[\mathscr{M}(D')\in\Omega]+\delta, \tag{6.7}$$

where ϵ is privacy budget which determines privacy level, and δ is a broken probability. The lower ϵ and δ represent the higher privacy level.

Definition 6.6. (Sensitivity) [44] For a $f_w : D\to R$, and neighboring datasets D and D', the sensitivity of f_w is defined as

$$\Delta f_w=\max_{D,D'}||f_w(D)-f_w(D')||. \tag{6.8}$$

Sensitivity measures the maximal difference between neighboring datasets.

Definition 6.7. (Gaussian mechanism) [48] Gaussian mechanism adds zero-mean Gaussian noise with variance $\Delta f_w^2 \sigma^2$ in each coordinate of the output $f_w(D)$,

$$\mathscr{M}(D) = f_w(D) + \mathscr{N}(0, \Delta f_w^2 \sigma^2), \tag{6.9}$$

where σ denotes the noise scale. It satisfies (ϵ, δ)-differential privacy if $\epsilon \in [0,1]$, $\delta \geq c\Delta f_w/\epsilon$, and $c^2 \geq 2ln(1.25/\delta)$.

DP-SGD [1] (see Algorithm 1 below) is an optimized SGD algorithm with two key steps to ensure that the gradients are updated in a differentially private way. Firstly, at iteration l, we denote the original gradients as g^l, the gradient after clipping as $\bar{g}^l$, and the gradient after clipping and adding noise as $\hat{g}^l$. The first step is that a clipping bound C is used on the l_2 norm of the gradient updates $g^l(x_i)$ with the batch size b (Line 5). The second step aggregates clipped gradients $\bar{g}^l$, and adds Gaussian noise $\mathscr{N}(0, \sigma^2 C^2)$ to the aggregation (Line 8). Since each update of $\hat{g}^l$ is differentially private, the final model parameters are also differentially private in terms of the composition property of differential privacy. The privacy leakage of DP-SGD is measured by (ϵ, δ), that is, the bound of privacy loss ϵ under a certain probability δ.

Algorithm 8 DP-SGD

Input: Dataset D, loss function $\mathscr{L}_D(w)$, learning rate r, batch size b, iteration times L, noise scale σ, clipping bound C

Initialize: Model parameters w^0

```
1: for $l \in L$ do
2:     Randomly sample a batch of data points $B^l(|B^l| = b)$ from $D$
3:     for $x_i \in B^l$ do
4:         $g^l(x_i) = \nabla_w^l \mathscr{L}(w^l, x_i)$
5:         $\bar{g}^l(x_i) = g^l(x_i) \times \min(1, \frac{C}{|g^l(x_i)|})$
6:     end for
7:     $\hat{g}^l = \frac{1}{b}\sum_i(\bar{g}^l(x_i) + \mathscr{N}(0, \sigma^2 C^2))$
8:     $w^{l+1} = w^l - r\hat{g}^l$
9: end for
```

Output: Model w^l and accumulated privacy cost (ϵ, δ)

6.2.4 Early Stopping Criteria

In the optimization process, high-capacity models tend to overfit. During the entire optimization process, when the loss on the training set decreases, the test loss saturates at a certain point and starts to increase again. Early stopping is widely used to solve this problem because it is easy to understand and implement, and in many cases is reported to be superior to regularization methods [87, 96, 112]. The gold standard for early stopping is to monitor the loss of the validation set. More specific, the continuous estimation of the generalization performance is observed in the validation set, the optimizer will stop when the generalization performance drops again. Different from traditional stopping criteria, we observe the discrimination level in the validation set and the optimizer will stop when the discrimination level is at a lower level after the training is convergent.

6.3 Preliminary Studies

In this section, we present some preliminary studies on bank deposit prediction and image classification tasks to show that DP-SGD results in decreased stability during the training process. Two main operations in DP-SGD, gradient clipping and noise addition protect the privacy of training data from the training process in deep neural networks. However, gradient clipping and noise addition also affect the direction of gradient update, resulting in decreased stability during the training process. The decreased stability brings more variations during the training, resulting in great variations in discrimination levels.

6.3.1 Preliminary Experiments

6.3.1.1 Bank Deposit Prediction: Gender and Bank Data

Dataset The Bank dataset [1] contains a total of 41,184 samples with 20 attributes and a binary label. The label indicates whether the client has subscribed to a term deposit or not. We consider 'Age' as the protected attribute. We sample a balanced Bank dataset with 30,000 samples (15,000 samples in each group).

[1] https://archive.ics.uci.edu/ml/datasets/bank+marketing

Model We use a logistic regression model with regularization parameter 0.01. We use SGD learning rate $r = 0.02$; and DP-SGD learning rate $r = 0.02$; batch size $b = 256$; the number of training epochs $T = 40$; clipping bound $C = 0.3$; and noise scale $\sigma = 1$.

6.3.1.2 Image Classification: Gender and Facial Images

Dataset The task is to make gender classification on the UTK Face dataset, which is a large-scale face dataset with a long age span [150]. We consider “Race" as the sensitive attribute, with white as the protected group and other races as the unprotected group, sampling 10,000 images from both groups. Before the model is applied, images are cropped to $200 * 200$ pixels.

Model We use a ResNet18 model [63] with 11M parameters pre-trained on ImageNet; SGD learning rate $r = 0.001$; DP-SGD learning rate $r = 0.01$; batch size $b = 128$; the number of training epochs $T = 100$; clipping bound $C = 1$; and noise scale $\sigma = 1$.

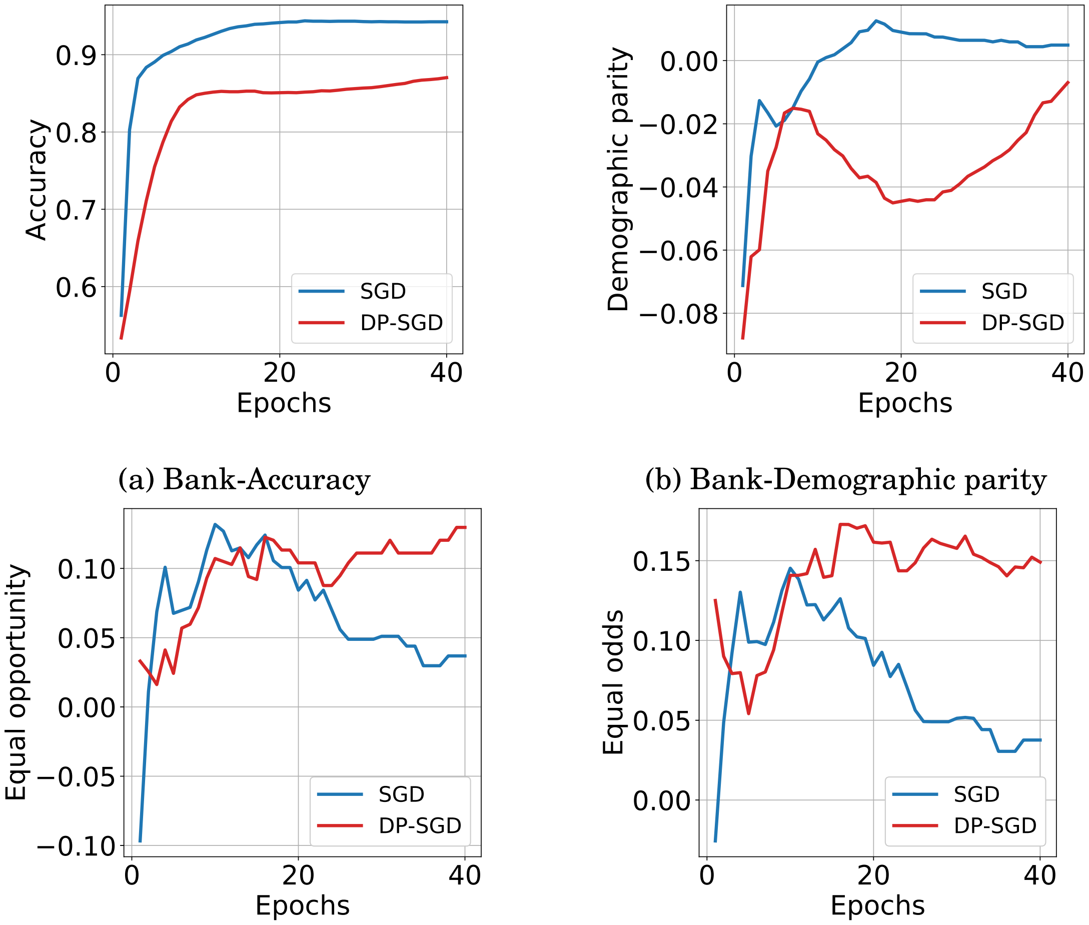
6.3.2 Results

Figures 6.2 and 6.3 show the impact of SGD and DP-SGD on model accuracy and fairness in terms of demographic parity, equal opportunity and equal odds on the Health and UTK Face datasets. Some observations can be seen based on the preliminary experiments:

1) Accuracy increases with growing training epochs in SGD and DP-SGD, and generally SGD has a faster increasing rate than DP-SGD and finally achieves higher accuracy.

2) DP-SGD training process is less stable than SGD’s because of gradient clipping and noise addition.

3) Even when the training converges, there are still small variations in accuracy, which lead to great variations in discrimination levels. Furthermore, it is noted that the discrimination level is at a low level in some epochs. This is because the discrimination level is more sensitive than accuracy. For example, one data point in the protected group is classified as label “1” in the previous epoch, then classified as label “1” in the current epoch. According to Definition 7, the effect is twofold, because positive predictions in the



(a) Bank-Accuracy

(b) Bank-Demographic parity

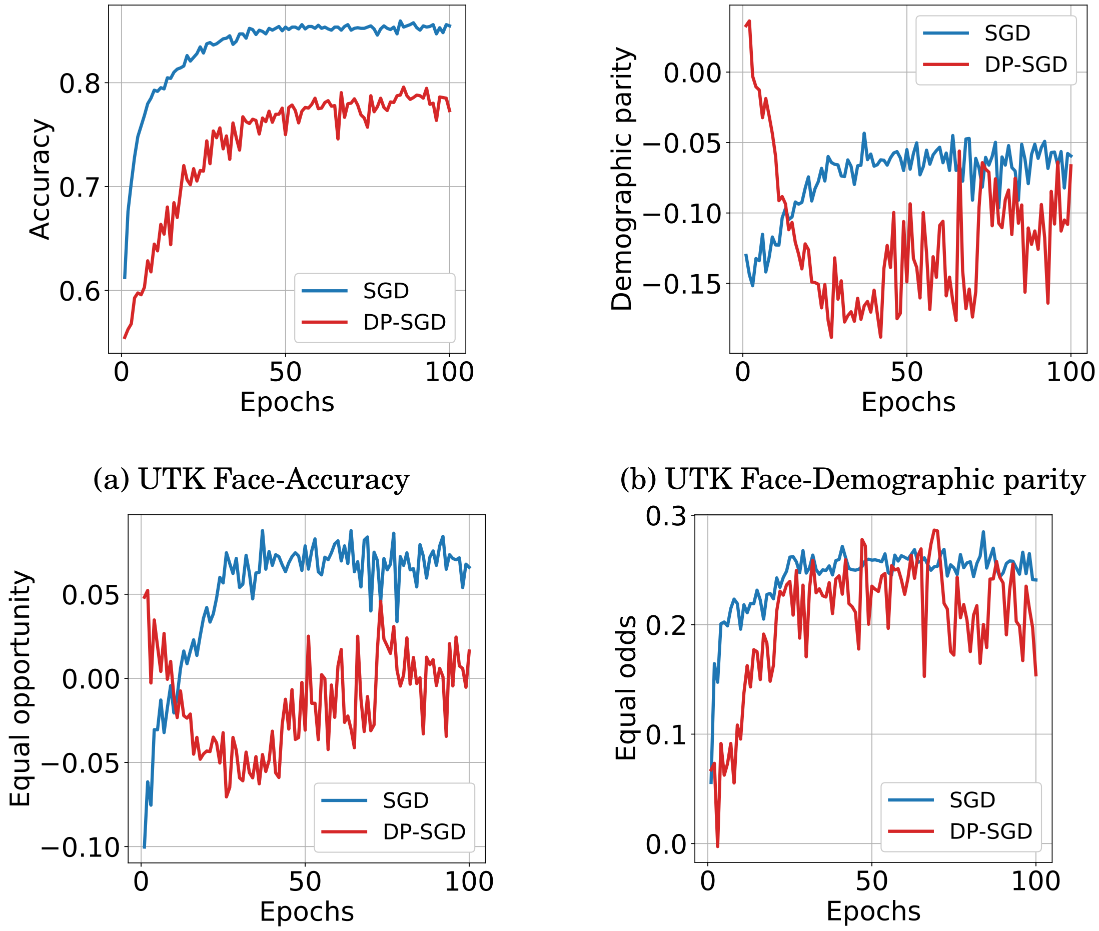
(c) Bank-Equal opportunity

(d) Bank-Equal odds

Figure 6.2: Training with SGD (Blue) and DP-SGD (Red) in the Bank dataset.

sensitive group are subtracted by 1, meanwhile, negative predictions in the unprotected group are increased by 1.

4) Variations are much more obvious in the UTK Face dataset than in the Health dataset. This is because the model used in the UTK Face dataset is much more complex (with 11M parameters) than the linear regression model used in the Health dataset. The more parameters the model has, the more dimensions the gradient has. Gradient clipping and noise addition have a greater impact on the gradients.



(a) UTK Face-Accuracy

(b) UTK Face-Demographic parity

(c) UTK Face-Equal opportunity

(d) UTK Face-Equal odds

Figure 6.3: Training with SGD (Blue) and DP-SGD (Red) in UTK Face dataset. Figure 3 shows that, with DP-SGD, more variations appear during the training in terms of accuracy and discrimination levels.

6.4 Early Stopping Criteria Methods

Based on the observation in the last section, DP-SGD brings more variations during the training, resulting in great variations in discrimination levels. In this section, we design two stopping criteria that can be used to halt training at the required trade-off between accuracy, privacy and fairness. Conventionally, the training for a deep neural network with SGD usually stops once the model converges. Yet, motivated by the evidence in all these analyses, it is clear that training could be stopped at a fluctuation that provides a desirable balance between accuracy, privacy, and fairness. The problem setting follows, and then each stopping criteria is outlined in turn.

Consider the training procedure in two stages: the first stage where accuracy increases rapidly over n epochs, and the second stage where accuracy stabilizes from epoch n to the end. The early stopping criteria are implemented during the second stage. In this way, the stopping point does not affect model accuracy much, and only focuses on how to stop the training when the discrimination level is lower. Let A be the accuracy of the training algorithm, where A^t_{val} denote the model's accuracy measured after the epoch t for the validation set.

Implementing the two criteria in the training process works as follows: (1) Train only on the training set and evaluate the per-example error on the validation set in every epoch. (2) Stop training as soon as the stopping criterion is satisfied. (3) Use the network weights from the previous step as the model. (4) Use the model on the test set to get the final results. This approach uses the validation set to anticipate the behavior on the test set, assuming that the trend of discrimination level will be similar for both sets. Algorithm 1 shows the implementation of two stopping criteria.

6.4.1 Stopping Criterion 1

The key idea of stopping criterion 1 is to stop the training when accuracy has a relative increase and discrimination level has a relative decrease at epoch t after training is convergent. Firstly, the value Γ^t_{max} is defined to be the highest discrimination in the validation set obtained from epoch n up to epoch $(n+k)$:

$$\Gamma^t_{max} = \max_{n \le t \le n+k} \Gamma^t_{val}. \tag{6.10}$$

The relative decrease of discrimination level $\Delta\Gamma$ in two epochs n and $(n+t)$ is defined as

$$\Delta\Gamma^t_{val} = \frac{\Gamma^t_{max}}{\Gamma^t_{val}} - 1. \tag{6.11}$$

The value A^t_{min} is defined to be the lowest accuracy in the validation set obtained from epoch n up to epoch $(n+k)$:

$$A^t_{min} = \min_{n \le t \le n+k} \Gamma^t_{val}. \tag{6.12}$$

The relative increase of accuracy ΔA in two epochs n and $(n+t)$ is defined as

$$\Delta A_{val}^{t} = \frac{A_{val}^{t}}{A_{min}^{t}} - 1 \tag{6.13}$$

Now we define $FA(t)$ as the sum of the relative decrease of discrimination level $\Delta\Gamma$ and the relative increase of accuracy ΔA at the training epoch t as

$$FA(t) = \alpha\Delta\Gamma_{val}^{t} + \Delta A_{val}^{t}, \tag{6.14}$$

where α is a parameter to control the balance between accuracy and discrimination level. FA can measure the balance between the discrimination level under certain fairness metrics and accuracy. A higher FA situation is that the discrimination level decreases at an epoch, while accuracy does not reduce much. A high FA is one obvious candidate reason to stop training because it indicates further training consumes more privacy cost and does not benefit model fairness and accuracy. This leads us to the first class of stopping criteria: stop as soon as the FA exceeds a certain threshold:

$$FA(t) \geq \beta_1, \tag{6.15}$$

where β_1 is the threshold.

6.4.2 Stopping Criterion 2

The second stopping criteria rely on the sign of changes in the discrimination level. This criterion says "stop when the discrimination level increases in β_2 successive strips:

$$\Gamma_{val}^{t-\beta_2} \geq \ldots \geq \Gamma_{val}^{t-1} \geq \Gamma_{val}^{t}. \tag{6.16}$$

The idea behind this definition is that when the discrimination level in the validation set has decreased not only once, but during β_2 consecutive strips, we assume that such a consecutive decrease indicates a local minimum of discrimination level. Note that if multiple epochs meet stop criterion 2, the training will stop at the first epoch that meets stop criterion 2.

Algorithm 9 Balance privacy, fairness and accuracy with early stopping criteria

```
Input: Training set $D_{tra}$, validation set $D_{val}$ loss function $\mathscr{L}(w)$, learning rate $r$, batch size $b$, parameters $\alpha, \beta_1, \beta_2$
Initialize: Model parameters $w^0$
Train the model on the $D_{tra}$, and evaluate the error on the $D_{val}$ with DP-SGD;
Find the epoch $n$ when accuracy stabilizes on the $D_{val}$;
Calculate the discrimination level $\Gamma$ in terms of fairness metrics and accuracy $A$ since epoch $n$;
—– Stopping criterion 1 —-
for $t \in [n, n+k]$ do
    Calculate $\Delta\Gamma^t_{val}$ and $\Delta A^t_{val}$
    if $FA(t) \geq \beta_1$ then
        Stop the training at the epoch $t$
    end if
end for
—– Stopping criterion 2 —-
for $t \in [n, n+k]$ do
    if $\Gamma^{t-\beta_2}_{val} \geq ... \geq \Gamma^{t-1}_{val} \geq \Gamma^t_{val}$ then
        Stop the training at the epoch $t$
    end if
end for
Output: Accuracy, discrimination level and accumulated privacy cost $(\epsilon, \delta)$
```

6.5 Theoretical Analysis

6.5.1 The Impact of DP-SGD on Gradient

In this section, we conduct an analysis of DP-SGD's impact on gradients w.r.t. each group. The analysis shows that gradient clipping leads to a disparate impact on model fairness. We analyze from the perspective of a single batch, where the utility loss is measured by the expected error of the estimated private gradient w.r.t. each group. Our analysis follows [7], which investigated bias-variance trade-off due to clipping. Let B^t be a batch of data points $x_i, ..., x_b$ at the epoch t. Each x_i corresponds to a data point and generates a gradient $g^t(x_i)$. The goal is to estimate the average gradient $\hat{g}^t$ from B^t in a differentially private way while minimizing the objective function.

Theorem 6.1. *Let original gradients at epoch t denote as g^t, and the gradient after clipping and noise as $\hat{g}^t$. The expected error of the estimated private gradient for a group* *z satisfies*

$$\mathbb{E}\left|\hat{g}_z^t - g_z^t\right| \leq \frac{1}{b_z}\frac{C}{\epsilon} + \frac{1}{b_z}\sum_i^{m_z}\left(\left|g_z^t(x_i)\right| - C\right), \tag{6.17}$$

where b_z is the size of a group z in a batch, and $m_z = |i : |g_z^t(x_i)| > C|$ is the number of data points that are clipped in the group z.

Proof. First, the original gradients g^t, the gradient after clipping $\bar{g}^t$, and the gradient after clipping and noise $\hat{g}^t$ are presented as following

$$g^t = \frac{1}{b}\sum_{i=1}^{b} g^t(x_i), \quad \bar{g}^t = \frac{1}{b}\sum_{i=1}^{b} \bar{g}^t(x_i), \tag{6.18}$$

$$\hat{g}^t = \frac{1}{b}\sum_{i=1}^{b}(\bar{g}^t(x_i) + \mathcal{N}(0, \sigma^2 C^2)). \tag{6.19}$$

The gradient after clipping and noise for the same batch of data points for a group z is denoted as, $\hat{g}_z^t = \frac{1}{b_z}(\sum_{i=1}^{b_z} \bar{g}_z^t + \mathcal{N}(0, \sigma^2 C^2))$ The expected error of the estimated private gradient of a group z can be decoupled as

$$\begin{aligned}
\mathbb{E}\left|\hat{g}_z^t - g_z^t\right| &\leq \mathbb{E}\left|\hat{g}_z^t - \bar{g}_z^t\right| + \left|\bar{g}_z^t - g_z^t\right| \\
&\leq \frac{1}{b_z}\frac{C}{\epsilon} + \frac{1}{b_z}\sum_i^{b_z}\max\left(0, \left|g_z^t(x_i)\right| - C\right) \\
&\leq \frac{1}{b_z}\frac{C}{\epsilon} + \frac{1}{b_z}\sum_i^{m_z}\left(\left|g_z^t(x_i)\right| - C\right).
\end{aligned} \tag{6.20}$$

■

Result 6.1. *The expected error of the estimated private gradient of a group z can be decoupled as bias and variance.*

From Inequation (6.20), we know that the utility loss of the group z, measured by the expected error of the estimated private gradient, is limited by two terms: the variance term and the bias term. The variance term, which is related to the clipping bound C and the privacy budget ϵ, produces the utility loss due to random noise added to the gradients. Bias is related to the clipping bound C and the size of gradients, which produces the utility loss by clipping gradients. Note that if C is small, the variance will be close to zero, but the bias will be large, making gradient less informative. Conversely, if C is large, the bias will be small, but the variance will be great.

Result 6.2. *The clipping bound could lead to different discrimination levels for each group when training neural networks with DP-SGD.*

Note that the bias in Inequation (6.20) is different for each group due to gradient distribution and the group size: these are the two factors that decide bias. Beginning with gradient distribution, the bias due to clipping is greater in the group with larger gradients. In SGD, the larger gradients make more of a contribution to the average gradient $\hat{g}^t$ prior to clipping, but this is not so with DP-SGD. After clipping, the larger gradients have lost more information, so the direction of the gradient update changes and is closer to the group with the smaller gradients. Second, the group with a larger group size makes more of a contribution in the average gradient $\hat{g}^t$ than the group with a smaller group size both before and after clipping. Hence, gradient clipping has a disparate impact on the gradients for each group and, in turn, the model accuracy. Since fairness metrics, such as demographic parity, and equal odds, are always related to predicted results from the model, disparate impact on model accuracy leads to disparate impact on model fairness.

6.5.2 The Impact of DP-SGD on the Training Process

Theorem 6.2. *Let $Var(\hat{g}^t)$ denote the variance of gradients at the epoch t in DP-SGD and $Var(g^t)$ denote the variance of gradients at the epoch t in SGD, and we have $Var(\hat{g}^t) \geq Var(g^t)$.*

Proof. To prove Theorem 6.2, we need to calculate the variance of gradients $Var(\hat{g}^t)$ in DP-SGD and the variance of gradients $Var(g^t)$ in SGD. The average of gradients w.r.t all data points is calculated as $g_{ave} = \sum_{i=1}^{N} g(x_i)$. The variance of gradients in SGD $Var(g^t)$ comes from sampling, and can be calculated as

$$Var(g^t) = \frac{1}{T}\sum_{t=1}^{T}(g^t - g_{ave})^2. \tag{6.21}$$

According to Expression in (6.21), gradients in DP-SGD are affected by gradient clipping and noise addition. The variance of gradients in DP-SGD $Var(\hat{g}^t)$ can be calculated as

$$Var(\hat{g^t}) = \frac{1}{T}\sum_{t=1}^{T}[(\hat{g}^t - g^t) + (g^t - g_{ave})]^2. \tag{6.22}$$

Comparing Equations (6.21) and (6.22), we can conclude $Var(\hat{g}^t) \geq Var(g^t)$. ■

Result 6.3. *In neural networks, the variance of the gradient can be transferred, which means that the variance of the gradient is large, and the variance of the model training result is also large. This can be expressed as* $Var(g^t) \propto Var(w) \propto Var(\hat{y}) \propto Var(\Gamma)$.

The variance of the gradients can transfer in model predictions as well as discrimination levels. The variance of the gradient $Var(g^t)$ is proportional to the variance of the model parameters $Var(w)$. This is because model parameters are updated by gradients directly according to Equation (1). The variance of the model parameters $Var(g^t)$ is proportional to the variance of the model predictions $Var(\hat{y})$. This is because the model predictions depend on the model parameters, and the relation can be simplified as $\hat{y} = f_w(x)$. The variance of model predictions $Var(\hat{y})$ is proportional to the variance of the discrimination level $Var(\Gamma)$. This is because the discrimination level $\Gamma(D, f)$ depends on model predictions. Hence, variance can be transitive from gradients to model predictions and discrimination levels. Combining with Theorem 2, DP-SGD results in a larger variance in accuracy and discrimination during the training.

6.5.3 The Changes of the Discrimination Level in the Validation Set and Test Set

Theorem 6.3. *Suppose that the validation set D_{val} and test set D_{test} come from the same data distribution $\mathscr{P}$, the changes in the discrimination level of the validation set $\Gamma(D_{val}, f_w)$ and the test set $\Gamma(D_{test}, f_w)$ are similar.*

Proof. Generally, the discrimination comes from the model and data [127], and thus we analyze the discrimination in the validation set D_{val} and test set D_{test} from the model and data. Firstly, the discrimination level in the metric of demographic parity in the dataset D can be expressed as follows

$$
\begin{aligned}
(6.23)\qquad & \Gamma^{DP}(D, f_w) \\
&= |\gamma_0(\hat{y}_D, f_w) - \gamma_1(\hat{y}_D, f_w)| \\
&= |Pr(f_w(x_i) = 1|z = 0) - Pr(f_w(x_i) = 1|z = 1|)| \\
&= |\frac{\sum_{i=1}^{N} f_w(x_i)(1 - z_i)}{\sum_{i=1}^{N}(1 - z_i)} - \frac{\sum_{i=1}^{N} f_w(x_i) z_i}{\sum_{i=1}^{N} z_i}|.
\end{aligned}
$$

Similarly, the discrimination level in the metric of Equal opportunity can be expressed as

$$\begin{aligned}\Gamma^{EO}(D,f_w) \\ = |\frac{\sum_{i=1}^{N} f_w(x_i)y_i(1-z_i)}{\sum_{i=1}^{N}(1-z_i)y_i} - \frac{\sum_{i=1}^{N} f_w(x_i)y_i z_i}{\sum_{i=1}^{N} z_i y_i}|.\end{aligned} \tag{6.24}$$

According to Equation (23) and (24), the discrimination level depends on the model f_w and the group size b_z. In our case, the same model is used in D_{val} and D_{test}, and thus the discrimination from the model should be similar. Also, D_{val} and D_{test} coming from the same data distribution $\mathscr{P}$ means that the group size b_z from each group should be approximately the same. Based on above, we conclude that the change of discrimination level should be similar in the validation set D_{val} and test set D_{test}. ■

6.6 Experiment

In this section, we first give more experimental results on the Medical prediction and sentiment analysis tasks to show the impact of DP-SGD on the training process. Second, we explore the impact of gradient clipping and noise addition on model accuracy and fairness. Third, we show the effects of two early stopping criteria in the image classification and sentiment analysis tasks under different noise scale σ, including accuracy, discrimination level, privacy cost, stopping epoch and parameter values. Finally, we present the positive prediction rate and true positive rate changes during the training for a better understanding of the changes in demographic parity and equal opportunity.

6.6.1 Experimental Setup

6.6.1.1 Medical Prediction: Age and Health Data

Dataset The task with the Health dataset [2] is to predict whether patients will be admitted to a hospital within the next year. We divide patients into two groups based on age (≤ 65 years and ≥65 years) and consider 'Age' as the sensitive attribute. We sample 10,000 data points from both two groups, so the Health dataset has 20,000 samples.

[2] https://foreverdata.org/1015/index.html

Model We use a logistic regression model with regularization parameter 0.01; SGD learning rate $r = 0.06$; DP-SGD learning rate $r = 0.2$; batch size $b = 256$; the number of training epochs $T = 40$; clipping bound $C = 0.4$; and noise scale $\sigma = 1$.

6.6.1.2 Sentiment analysis: Ethnicity and Movie Reviews

Dataset The last experiment is a sentiment analysis task using the IMDB movie review dataset [93], and the objective is to classify each review as positive or negative. The sensitive attribute is ethnicity, given by the method in [12, 13], where reviews are labeled as Standard American English (SAE) or Other American English (OAE). We sample 2,5000 tweets from each group split equally between positive and negative sentiments.

Model We use FastText model [71] with 2.5M parameters, and pretrained 100-dimensional GloVe embedding [110]. We use SGD learning rate $r = 0.9$; DP-SGD learning rate $r = 0.6$; batch size $b = 64$; the number of training epochs $T = 100$; clipping bound $C = 1$; and noise scale $\sigma = 1$.

6.6.1.3 Protocol

In the experiments, the learning rate is different for SGD and DP-SGD on the dataset. This is because the convergence rate of DP-SGD depends on the variance of stochastic gradients, bias introduced due to gradient clipping and variance due to noise addition [**?**]. The learning rate is tuned for the best performance by each model after binary searching. Parameter α is set to be 0.001 in Eq.(14). The range of β_1 is from 0.02 to 2, and β_2 is set to be 3.

In all settings, the dataset is split into 60% training data, and 20% validation data and 20% testing data. δ is set to as 10^{-5}. The implementation is based on PyTorch [109]. Experiments were carried out with NVIDIA GeForce GTX 1080 Ti, NVIDIA Quadro RTX 5000 and Intel Xeon Gold 6142.

6.6.1.4 Evaluation Metrics

We define a metric that is used to evaluate the trade-off between accuracy, fairness and privacy.

Definition 6.8. Let A^t_{test} and Γ^t_{test} denote the model accuracy and discrimination level measured after the epoch t for the test set. We define the trade-off between accuracy, fairness and privacy (AFP) as the linear combination of accuracy, discrimination and training epoch, which is expressed as follows

$$AFP = -\lambda_1 A^t_{test} + \lambda_2 \Gamma^t_{test} + \lambda_3 t \tag{6.25}$$

where λ_1, λ_2, and λ_3 are parameters that adjust the importance of accuracy, discrimination level and privacy cost. Since privacy cost is linearly related to training epoch t when noise scale σ and clipping bound C are fixed in DP-SGD, we use epoch t to measure privacy cost. A smaller AFP indicates a better trade-off, which means that the training stops at higher accuracy and a lower discrimination level as early as possible when the training is convergent. In the experiment, we set $\lambda_1 = 0.1$, $\lambda_2 = 1$ and $\lambda_3 = 0.01$.

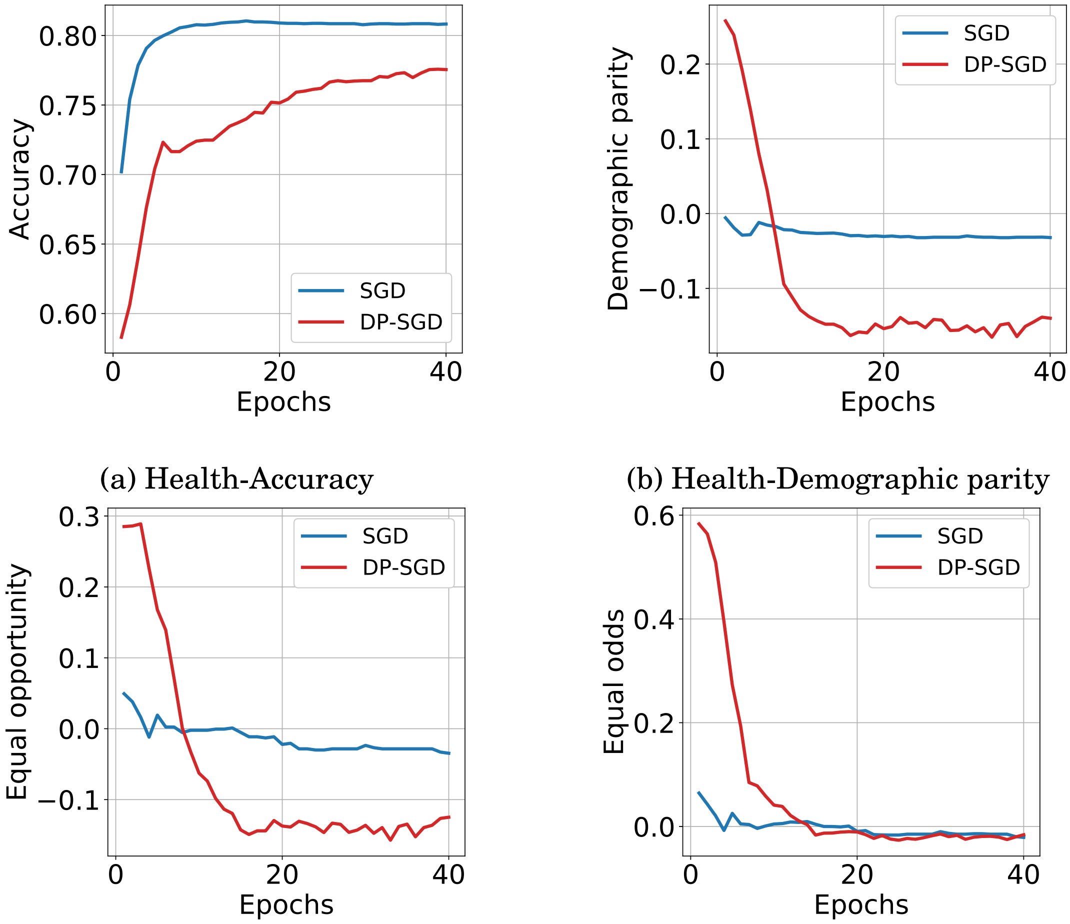
6.6.1.5 Baseline

Baseline 1: We set the baseline 1 that the epoch with the best validation accuracy is used as the epoch to stop training after the training is convergent. This is a commonly used method to implement traditional early stopping and is likely to achieve the best test accuracy. More specifically, the validation accuracy in the UTK Face dataset and the IMDB dataset stops increasing roughly after 70 epochs, and we choose the best validation accuracy from epoch 70 to epoch 100.

Baseline 2: We set the baseline 2 that the training stops at the 30-th epoch after the training is convergent. This is another commonly used method to implement traditional early stopping, that is, stopping the training at a certain epoch after the training is convergent. More specifically, since the training on the UTK Face dataset and the IMDB dataset has converged since epoch 70, we set the training of both datasets to stop at epoch 100.

6.6.2 Accuracy and Discrimination Level with SGD and DP-SGD

Figures 6.4 and 6.5 show the impact of SGD and DP-SGD on model accuracy and fairness in terms of three fairness metrics on the Health and IMDB datasets. As shown, accuracy again increases with more training epochs, and SGD's accuracy is higher than DP-SGD's. DP-SGD's training process is smoother than SGD's. Even when the



(a) Health-Accuracy

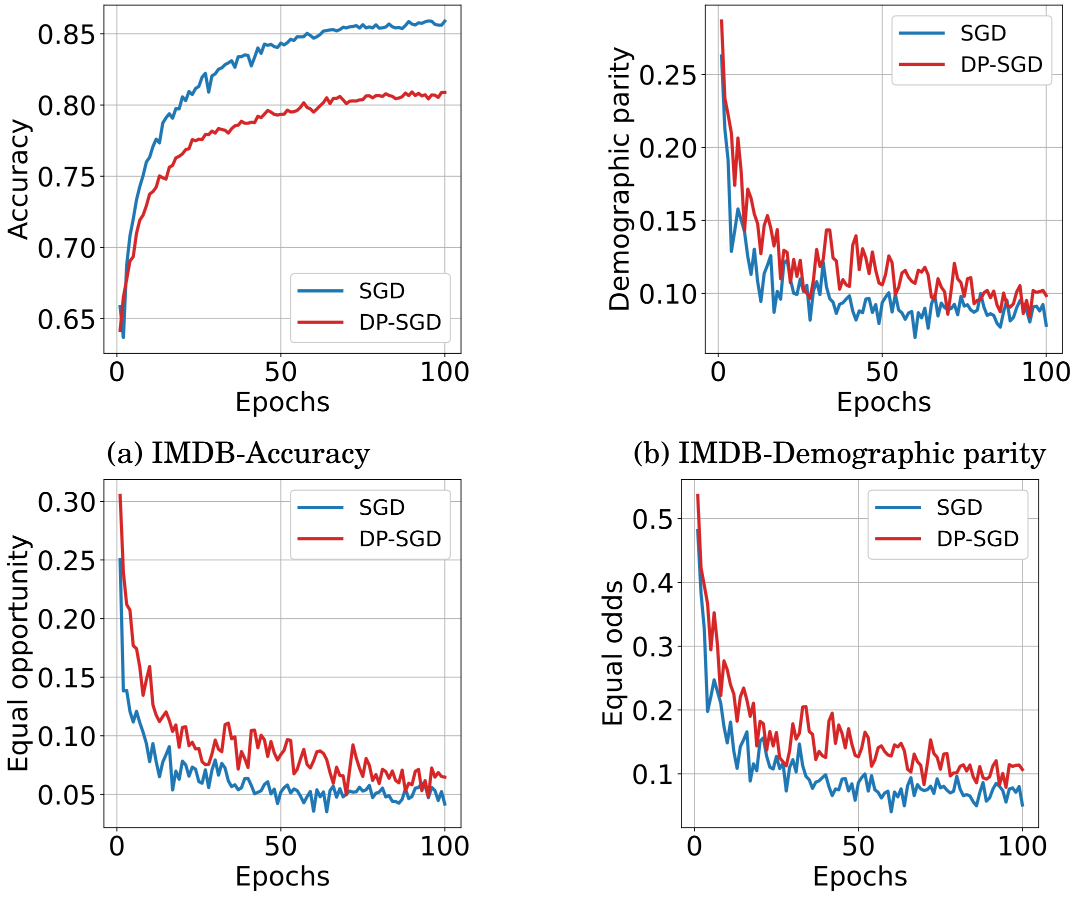
(b) Health-Demographic parity

(c) Health-Equal opportunity

(d) Health-Equal odds

Figure 6.4: Training with SGD (Blue) and DP-SGD (Red) in the Health dataset.

training converges, there are still small variations in accuracy, which lead to great variations in discrimination levels. All fairness metrics fluctuate during training, and hence the fairness levels are highly related to when training is halted. Observations from Figures 6.2-6.5 confirm our analysis in Section 6.4. That is gradient clipping and noise addition in DP-SGD lead to high variations during the training process. When the training is convergent, small variations in accuracy may lead to dramatic variations in discrimination levels.



(a) IMDB-Accuracy

(b) IMDB-Demographic parity

(c) IMDB-Equal opportunity

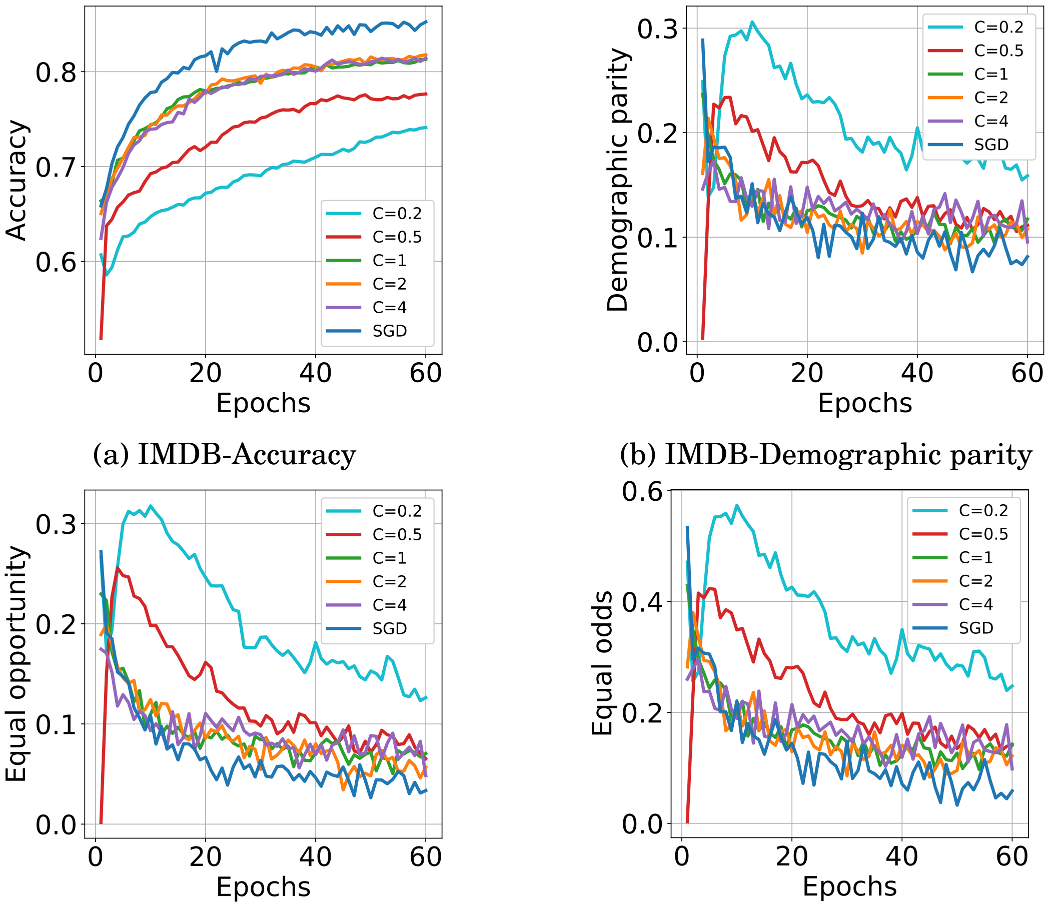
(d) IMDB-Equal odds

Figure 6.5: Training with SGD (Blue) and DP-SGD (Red) in the IMDB dataset. Figure 6.5 shows that, with DP-SGD, more variations appear during the training in terms of accuracy and discrimination levels.

6.6.3 Effect of Hyperparameters

We further conducted empirical analyses with the same three scenarios to examine the impact of clipping bound C and noise scale σ. The results for each case are similar. Therefore, we report the results of the sentiment analysis with the IMDB dataset here.

With the noise ($\sigma \cdot C$) fixed to 1, and the clipping bound C varied from $0.2, 0.5, 1, 2, 4$, Figure 6 shows that accuracy is better with a larger C until around 1, after which it does not change much. This is because a larger C clips less information up to a point (at 1) where no information is clipped at all. A smaller C, however, has a higher



(a) IMDB-Accuracy

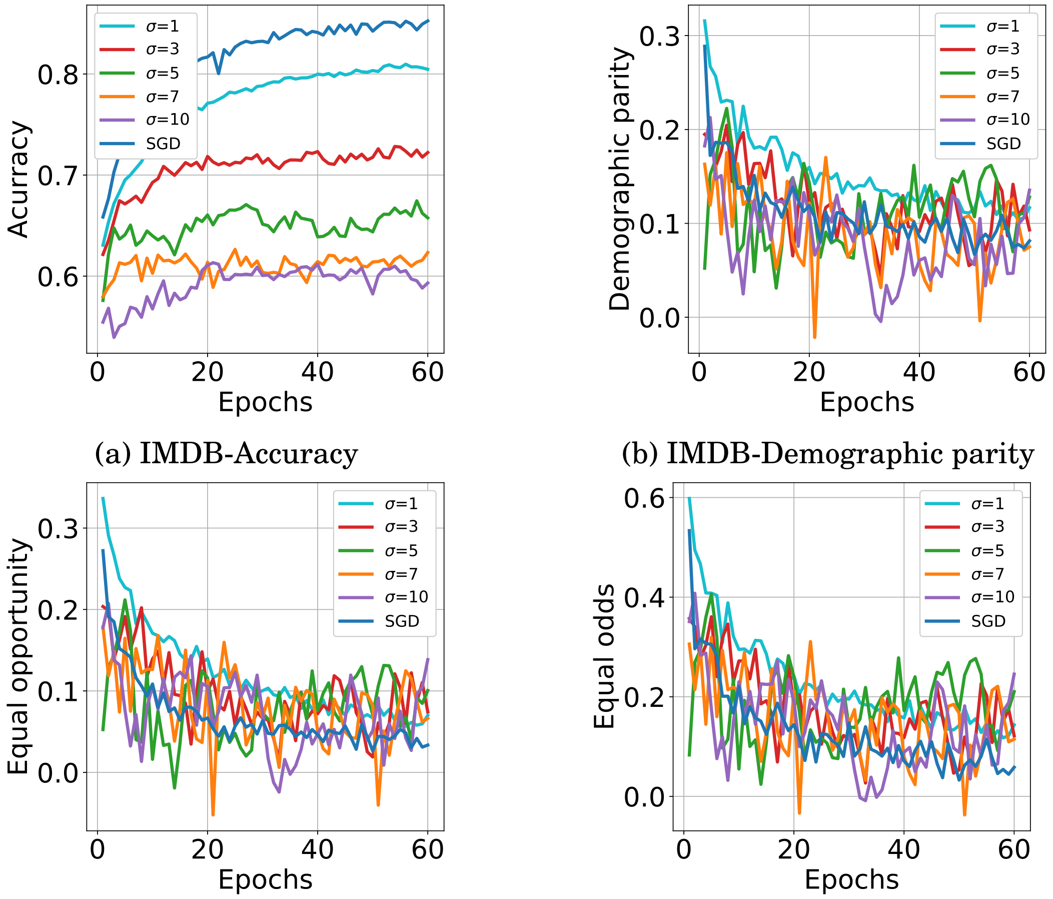
(b) IMDB-Demographic parity

(c) IMDB-Equal opportunity

(d) IMDB-Equal odds

Figure 6.6: Training with SGD (Dark blue) and DP-SGD (other colors; $\sigma \cdot C = 1$ and C varies) in the IMDB dataset. Figure 6.6 shows that clipping bound has an impact on discrimination levels.

disparate impact and, therefore, the final model has a high discrimination level. Figure 6.6 shows the reverse, with the clipping bound C fixed at 1, and the noise $(\sigma \cdot C)$ varied among $1, 3, 5, 7, 10$. Figure 6.7 shows the impact of noise level on SGD and DP-SGD. It is clear, and unsurprising, that the amount of noise added has a huge impact on accuracy. However, the same is not true of fairness. Noise does have some impact, but it is not decisive.



(a) IMDB-Accuracy

(b) IMDB-Demographic parity

(c) IMDB-Equal opportunity

(d) IMDB-Equal odds

Figure 6.7: Training with SGD (Dark blue) and DP-SGD (other colors; $C = 1$ and σ varies) in the IMDB dataset. Figure 7 shows that the noise scale does not produce a distinctive difference in discrimination levels.

6.6.4 The Effect of Stopping Criteria

With structured datasets, training tends to be smooth because the models are usually simple. Thus, stopping criteria are more useful in deep neural networks, and we conducted two stopping criteria on unstructured datasets. We conduct two early stopping criteria (SC 1 and SC 2) in different privacy levels, including $\sigma = 1$, $\sigma = 1.5$ and $\sigma = 2$. β (β_1 in the SC 1, β_2 in the SC 2) is the parameter in the early stopping criteria, and the value of β is chosen by some simple attempts. For example, when the value of β is large, the stopping criteria may not be able to stop the training. In this case, the value of β

UTK Face dataset												
Metrics	Demographic parity				Equal opportunity				Equal odds			
	Acc	Dis	Epoch	AFP	Acc	Dis	Epoch	AFP	Acc	Dis	Epoch	AFP
SC 1, $\sigma = 1$	0.7732	**0.0915**	89	0.9041	0.7494	**0.0224**	74	**0.6874**	0.7797	**0.1760**	**90**	**0.9980**
SC 2, $\sigma = 1$	0.7708	0.1169	**79**	**0.8298**	0.7708	0.0227	**79**	0.7356	0.7797	**0.1760**	**90**	**0.9980**
Baseline 1, $\sigma = 1$	**0.7880**	0.1686	94	1.0308	**0.7880**	0.0242	94	0.8854	**0.7880**	0.2964	94	1.1576
Baseline 2, $\sigma = 1$	0.7851	0.1245	100	1.045	0.7851	0.0248	100	0.9462	0.7851	0.2624	100	1.1838
SC 1, $\sigma = 1.5$	**0.7821**	0.1062	81	0.8379	0.7791	**0.0018**	85	0.7738	0.7833	**0.1633**	**91**	0.9953
SC 2, $\sigma = 1.5$	0.7791	**0.0832**	**74**	**0.7452**	0.7750	0.0070	**79**	**0.7190**	**0.7851**	0.1968	96	1.0188
Baseline 1, $\sigma = 1.5$	0.7755	0.1349	80	0.8573	0.7755	0.0051	80	0.7275	0.7755	0.2576	80	**0.9800**
Baseline 2, $\sigma = 1.5$	0.7815	0.1130	100	1.0348	**0.7815**	0.0107	100	0.9325	0.7815	0.2270	100	1.1488
SC 1, $\sigma = 2$	0.7696	0.0934	93	0.9464	0.7607	0.0221	**76**	**0.7060**	0.7535	0.1849	**77**	**0.8795**
SC 2, $\sigma = 2$	0.7519	**0.0758**	**79**	**0.7906**	0.7696	**0.0048**	93	0.8578	0.7517	**0.1449**	81	0.8797
Baseline 1, $\sigma = 2$	**0.7726**	0.1158	97	1.008	**0.7726**	0.0085	97	0.9012	**0.7726**	0.2456	97	1.1383
Baseline 2, $\sigma = 2$	0.7684	0.1391	100	1.062	0.7684	0.0271	100	0.9502	0.7684	0.2445	100	1.1676

Table 6.1: Accuracy, discrimination levels and training epoch for the two early stopping criteria with different parameters on the UTK Face dataset

could be adjusted to a smaller value.

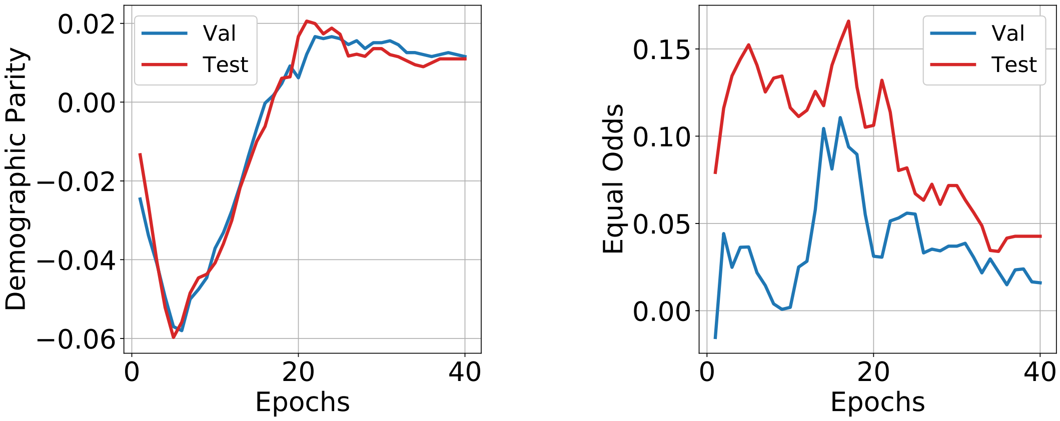
Tables 6.1 and 6.2 show the accuracy, discrimination level and stopping epoch for each of the two stopping criteria with the UTK Face and IMDB datasets. As shown, baselines present a higher AFP most of the time, compared with the training with proposed stopping criteria. This indicates that two stopping criteria can stop the training at a better trade-off between accuracy, fairness and privacy. This is because DP-SGD brings variations during the training, which leads to the possibility to obtain a model that has a small discrimination level. In many cases, stopping criteria help to reduce nearly half of the discrimination level. For example, the discrimination level can be reduced to 0.0694 (when $\sigma = 2$ using SC 1) in the metric of demographic parity in the IMDB dataset, which is less than half of the discrimination level in baselines.

6.6.5 Discrimination Levels in Validation and Test Sets

To explain why stopping criteria is helpful to guide when to stop the training, we show how discrimination levels change in the validation and test sets of the sentiment analysis task with IMDB. Figures 6.8-6.11 shows the trend of discrimination levels in terms of demographic parity and equal opportunity on four datasets. The trend of discrimination levels is consistent in validation and test sets, even if a gap may exist in discrimination levels. The observation confirms the analysis in Section 6.3. Thus, the validation set can help us to signal the ideal time to stop the training with the test set.

IMDB dataset												
Metrics	Demographic parity				Equal opportunity				Equal odds			
	Acc	Dis	Epoch	AFP	Acc	Dis	Epoch	AFP	Acc	Dis	Epoch	AFP
SC 1, $\sigma = 1$	0.8064	**0.0821**	92	0.9214	0.8064	**0.0530**	92	0.8923	0.8064	**0.0827**	92	0.8920
SC 2, $\sigma = 1$	0.8050	0.0921	**77**	**0.7816**	0.8050	0.0612	**77**	**0.7507**	0.8079	0.1133	**81**	**0.8225**
Baseline 1, $\sigma = 1$	**0.8092**	0.0929	89	0.9019	**0.8092**	0.0581	89	0.8671	**0.8092**	0.1052	89	0.9142
Baseline 2, $\sigma = 1$	0.8088	0.0985	100	1.0176	0.8088	0.0658	100	0.9849	0.8088	0.1119	100	1.0310
SC 1, $\sigma = 1.5$	0.7792	**0.0782**	94	0.9402	0.7782	**0.0518**	85	0.8239	0.7782	**0.0819**	85	0.8545
SC 2, $\sigma = 1.5$	0.7788	0.0940	**77**	**0.7861**	0.7794	0.0639	**73**	**0.7159**	0.7788	0.1067	**77**	**0.7996**
Baseline 1, $\sigma = 1.5$	0.7777	0.1030	95	0.9752	0.7777	0.0672	95	0.9394	0.7777	0.1258	95	0.9980
Baseline 2, $\sigma = 1.5$	**0.7812**	0.0945	100	1.0163	**0.7812**	0.0608	100	0.9826	**0.7812**	0.1019	100	1.0476
SC 1, $\sigma = 2$	**0.7609**	**0.0720**	86	0.8559	**0.7609**	0.0392	86	0.8231	**0.7609**	**0.0694**	86	0.8533
SC 2, $\sigma = 2$	0.7604	0.0740	**81**	**0.8079**	0.7604	**0.0273**	**81**	**0.7612**	0.7604	0.0705	**81**	**0.8044**
Baseline 1, $\sigma = 2$	0.7592	0.1180	93	0.9720	0.7592	0.0949	93	0.9489	0.7592	0.1607	93	1.0147
Baseline 2, $\sigma = 2$	0.7590	0.1077	100	1.0318	0.7590	0.6288	100	0.9869	0.7590	0.1388	100	1.0629

Table 6.2: Accuracy, discrimination level and training epoch for the two early stopping criteria with different parameters on the IMDB dataset



(a) Bank-Demographic parity

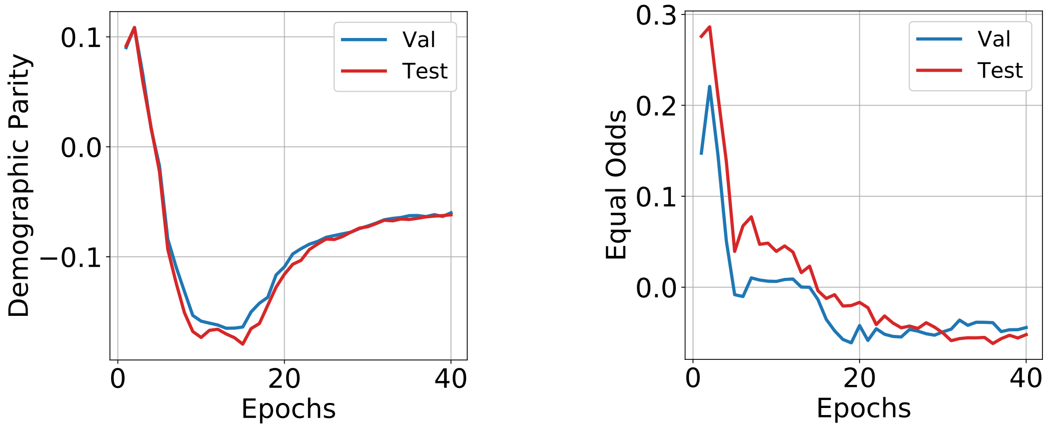
(b) Bank-Equal odds

Figure 6.8: Discrimination level of demographic parity (a) and equal odds (b) on validation set (Blue) and test set (Red) in the Bank dataset. Figure 6.8 shows the changes of discrimination level is similar in validation and test set in the Bank dataset.

6.6.6 Positive Prediction Rate and True Positive Rate

For a better understanding of how DP-SGD affects metrics of demographic parity and equal opportunity, we plot the positive prediction rate (PPR) and true positive rate (TPR) changes during training in each group. As Definitions 4 and 5 describe, demographic parity is the difference in the number of positive predictions between groups, and equal opportunity is the difference in TPR between groups.

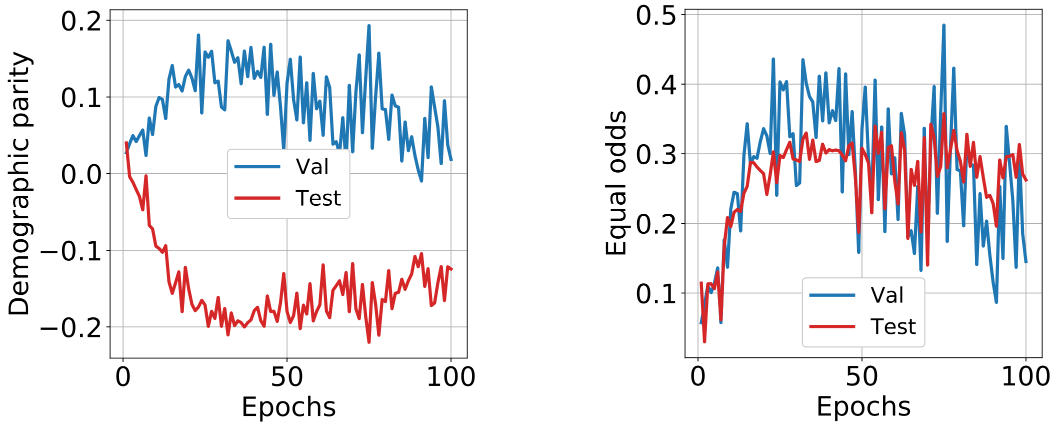
In Figure 6.12, the positive prediction rate and true positive rate are both higher



(a) Health-Demographic parity

(b) Health-Equal odds

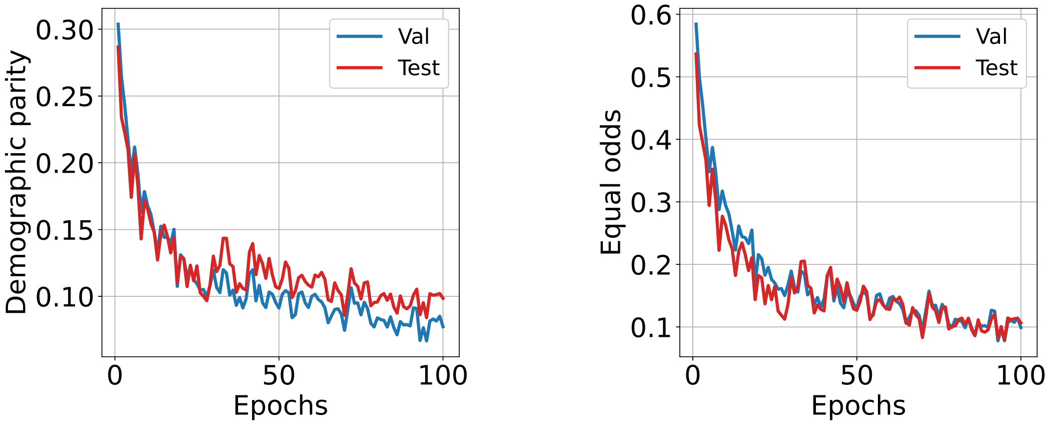
Figure 6.9: Discrimination level of demographic parity (a) and equal odds (b) on validation set (Blue) and test set (Red) in the Health dataset. Figure 6.9 shows the changes of discrimination level is similar in validation and test set in the Health dataset.



(a) UTK Face-Demographic parity

(b) UTK Face-Equal odds

Figure 6.10: Discrimination level of demographic parity (a) and equal odds (b) on validation set (Blue) and test set (Red) in the UTK Face dataset. Figure 6.10 shows the trend of discrimination level is similar in validation and test set.



(a) IMDB-Demographic parity

(b) IMDB-Equal odds

Figure 6.11: Discrimination levels on the validation set (Blue) and test set (Red) in the IMDB dataset. Figure 6.11 shows the trend of discrimination levels is similar in both sets.

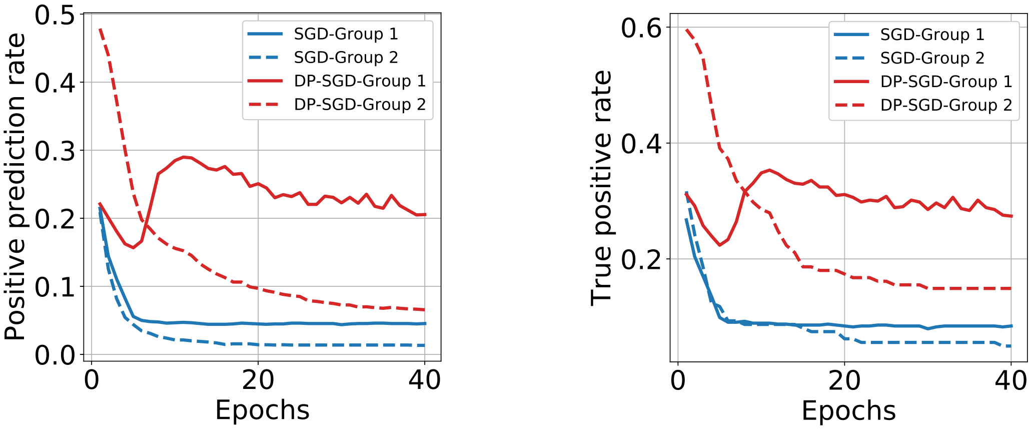
in Group 2 than Group 1, and DP-SGD has a bigger impact on Group 2 than Group 1. In Figures 6.13 and 6.14, we can see a great deal of variation in both rates. This is because deep neural network models typically deal with high-dimensional gradients, which means gradient clipping has a higher probability of affecting gradients during training. Even so, we could still see the trend of positive prediction rate and true positive rate for each group. Overall, we find that DP-SGD changes the positive prediction rate, and further affects discrimination level in demographic parity and equal odds. This indicates that gradient clipping affects positive predictions during the training.

6.6.7 Discussion and Summary

6.6.7.1 Discussion

As shown in Tables 6.1 and 6.2, when σ increases, stopping criteria are more likely to obtain the model with smaller discrimination. This is because a larger σ brings more noise during the training, and noise leads to variations. This provides the possibility that the model is updated at a low discrimination epoch.

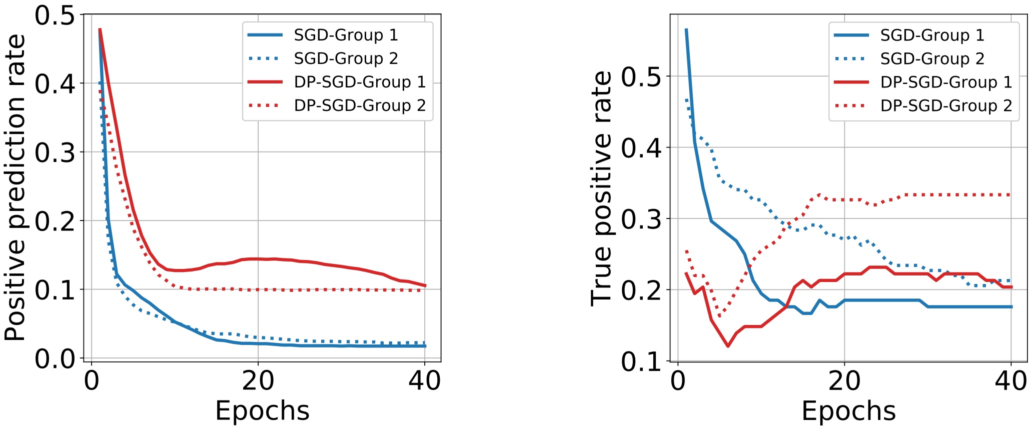
Two stopping criteria have some differences. β_1 is an important parameter to determine the trade-off. An appropriate β_1 in stopping criterion 1 can stop the training at a desirable trade-off. Since stopping criterion 1 considers the variation of accuracy and



(a) Health-PPR

(b) Health-TPR

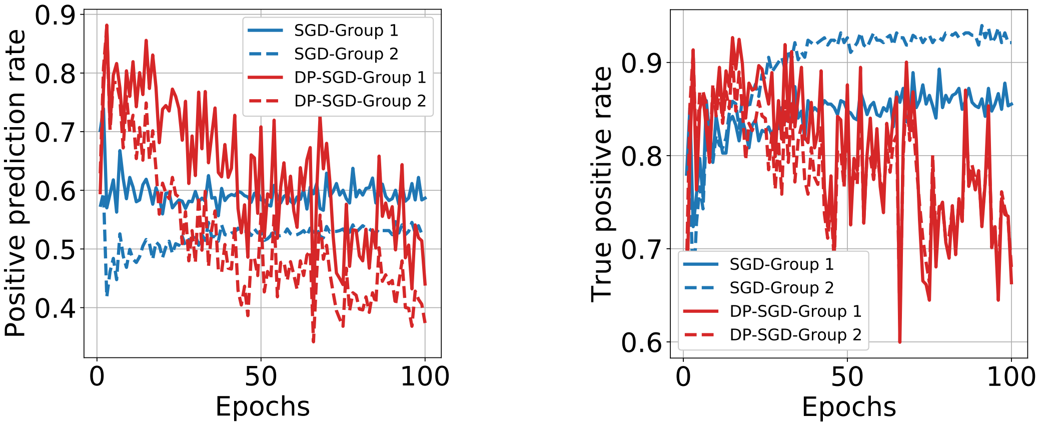
Figure 6.12: Training with SGD (Blue) and DP-SGD (Red) in the Health dataset. (a) Positive predication rate in Group 1 (Solid line) and Group 2 (Dotted line) (b) True positive rate in Group 1 (Solid line) and Group 2 (Dotted line). Figure 6.12 shows how PPR and TPR change in two groups in the Health dataset.



(a) Bank-PPR

(b) Bank-TPR

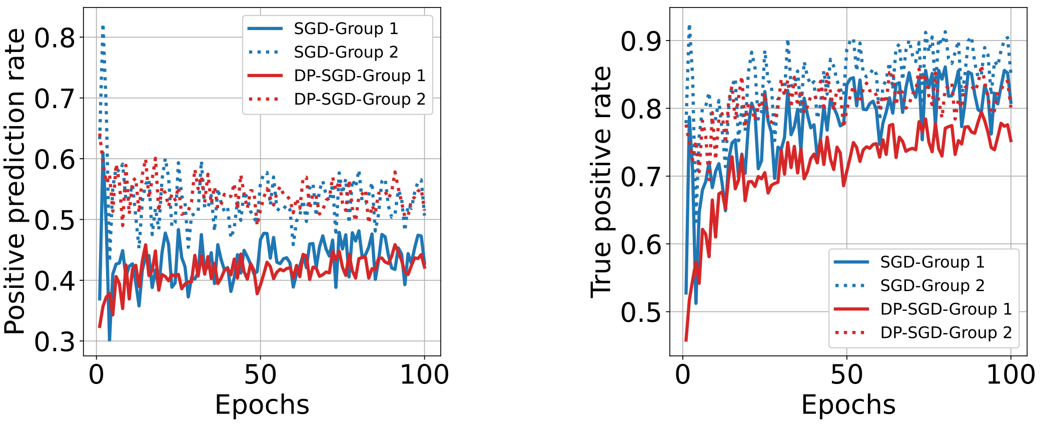
Figure 6.13: Training with SGD (Blue) and DP-SGD (Red) in the Bank dataset. (a) Positive prediction rate in the protected group (solid line) and unprotected group (Dotted line) (b) True positive rate in protected group (solid line) and unprotected group (Dotted line). Figure 6.13 shows how PPR and TPR change in two groups in the Bank dataset.



(a) IMDB-PPR

(b) IMBD-TPR

Figure 6.14: Training with SGD (Blue) and DP-SGD (Red) in the IMDB dataset. (a) Positive prediction rate in Group 1 (Solid line) and Group 2 (Dotted line) (b) True positive rate in Group 1 (Solid line) and Group 2 (Dotted line). Figure 6.14 shows how PPR and TPR change in two groups in the IMDB dataset.



(a) UTK Face-PPR

(b) UTK Face-TPR

Figure 6.15: Training with SGD (Blue) and DP-SGD (Red) in the UTK Face dataset. (a) Positive predication rate in Group 1 (Solid line) and Group 2 (Dotted line) (b) True positive rate in Group 1 (Solid line) and Group 2 (Dotted line). Figure 6.15 shows how PPR and TPR change for two groups in the UTK Face dataset.

discrimination level, it can avoid stopping the training at some epochs when accuracy is not good. Stopping criterion 2 is easy to implement and parameter β_2 is easier to select. However, stopping criterion 2 may not stop the training at the best trade-off.

6.6.7.2 Summary

From these experiments, we summarize some findings as follows. (1) When DP-SGD is used in deep neural network models, gradient clipping and noise addition will bring more variations in model accuracy and fairness. (2) Both theoretical analysis and experiments express a similar trend in validation and test sets in terms of demographic parity, equal opportunity and equal odds. (3) Stopping criteria helps to strike a better balance between accuracy, privacy and fairness. In terms of accuracy, the training is stopped when it is convergent, which means little accuracy is lost. In terms of privacy, stopping training early helps to save the privacy cost. In terms of fairness, the designed criteria provide two means of striking a desirable discrimination level.

6.7 Related work

6.7.1 Differentially Private Machine Learning

In recent years, many work have studied differential privacy in machine learning. In terms of the position where random noise is added in the training process, we classify existing research into four types: input perturbation, output perturbation, objective perturbation and gradient perturbation. Input perturbation means that individual data are randomly perturbed to some extent before they are handed over to the model for learning or analysis to prevent the model from acquiring real data [54]. Chaudhuri et al. provided output perturbation where noise is added to optimal parameters obtained by empirical risk minimization [24, 108], and an objective perturbation where noise is added to the objective function [25, 56]. Later, differentially private stochastic gradient, where noise is added to in the process of solving the optimal model parameters using the gradient descent method, and ensuring that the entire process meets differential privacy [1, 144]. In order to ensure the calculation efficiency of the algorithm, stochastic gradient descent or small batch gradient descent methods are often used in practical applications.

6.7.2 Fair Machine Learning

In recent years, a large number of researchers have been working on fair machine learning, we summarize the existing work on fair machine learning in the following three lines.

6.7.2.1 Pre-processing Methods

Pre-processing methods eliminate the discrimination by adjusting the training data, including suppression, reweighing and sampling to obtain fair datasets before training [16, 73]. Also, learning fair intermediate representations in pre-process phase has received much attention. [147] was the first to achieve fair machine learning by learning fair intermediate representations. The basic idea is that mapping the training data to a transformed space where as much useful information as possible is retained, but the dependencies between sensitive attributes and class labels are removed. Many researchers studied fair representation learning with different methods, such as adversary learning [52, 94, 124]. These methods are based on using a classifier to predict sensitive attributes as adversarial components. The advantage of pre-precessing methods is that these methods can apply to all algorithms and tasks.

6.7.2.2 In-processing Methods

In-processing methods avoid discrimination with the modifications in the model aspects, such as adding fair constraints or regularizers. [75] used regularizer term to penalize discrimination to enforce non-discrimination in the learning objective. [57, 145, 146] designed the convex fairness constraint to achieve fair classification. [4] reduced fair classification to a sequence of cost-sensitive classification problems, where fairness definitions are formalized as linear inequalities. [5] built a unified framework for designing optimal and fair decision trees for classification and regression decisions. [98] studied fair classifiers that were robust to perturbations, and formulated the problem as a min-max to minimize a distributionally robust training loss. The advantage of in-processing methods is that the level of fairness and accuracy can be controlled by parameters in the model.

6.7.2.3 Post-processing Methods

A third approach to achieving fairness is post-processing, where a learned classifier is modified to adjust the decisions to be non-discriminatory for different groups. [60] proposed an approach to use of post-processing to ensure fairness criteria of equal opportunity and equal odds.

6.7.3 Comparison with Other Work

Several papers have studied the privacy and fairness issues in machine learning simultaneously. [40, 139] proposed a differentially private fair algorithm in logistic regression. [67] studied fair learning under the constraint of differential privacy in the equalized odds. [22] studied privacy risks of group fairness through the lens of membership inference attacks. Different from these papers working on margin classifiers, our work studies the balance between fairness, privacy and accuracy in deep neural networks. And while, [9, 138] have considered the impact of DP-SGD on model accuracy. We first put it further, we conduct preliminary studies on the impact of DP-SGD on model fairness. We then analyze how to reach a better trade-off between accuracy, fairness and privacy with early stopping criteria.

6.8 Summary

The empirical and theoretical analyses show that DP-SGD makes the training less stable, and has a disparate impact on model fairness. Moreover, small variations in accuracy can lead to dramatic variations in the resulting model‚Äôs discrimination levels, and the variations are similar in the validation set and test set. Hence, we leverage these phenomena with two different criteria for stopping training as a solution for fine-tuning the trade-off between accuracy, fairness and privacy. The stopping criteria are easy to implement and can be used with any fairness metric.

CHAPTER 7

REVISITING MODEL FAIRNESS VIA ADVERSARIAL EXAMPLES

In the last main chapter, we have studied how privacy-preserving methods could affect model fairness. However, fairness could also be affected by adversarial examples. Existing research evaluates model fairness over limited observed data. In practice, however, factors such as maliciously crafted examples and naturally corrupted examples often appear in real-world data collection. This severely limits the reliability of bias removal methods, inhibits the fairness improvement in long-term learning systems, and probes to study accuracy-related robustness. In this chapter, we focus on the vulnerability of model fairness - how adversarial examples could skew model fairness. We propose a general adversarial fairness attack framework that can twist model bias through a small subset of adversarial examples.

7.1 Introduction

Model fairness has been studied extensively [31, 105], with researchers agreeing that it is is an essential in machine learning models. Many events have proven that machine learning algorithms can discriminate and do harm to the basic rights of human beings. For example, Vigdor et al. [131] reported gender bias in Apple card's credit ranking, while

GPT-3, a state-of-the-art language model has been found to capture persistent Muslim-violence bias [2]. Most research on fair machine learning has focused on defining what is fair in machine learning, including individual fairness [46] and group fairness [15, 145], and explored the methods to remove discrimination before learning [117, 124, 149], during learning [4, 146] and after learning [91]. However, the literature has largely ignored the susceptibility of assessing model fairness. The fairness of classifiers is often evaluated on sampled datasets, and can be unreliable for various reasons, including biased examples, missing and/or noisy attributes [69, 84, 98]. Assessing model fairness is the key to determining the effectiveness of bias removal methods, as well as the key to improving model fairness in dynamic learning systems [37, 89].

In this work, we are the first to focus on an important yet under-studied aspect of the fairness problem: the vulnerability of model fairness to adversarial attacks. Learning models are proven vulnerable to adversarial examples [58, 102, 122]. A small and imperceptible perturbation on data examples is sufficient to fool state-of-the-art classifiers, resulting in incorrect classification. However, research in adversarial machine learning has mostly focused on accuracy. We argue that, like accuracy, model fairness can also be targeted by malicious adversaries. Intuitively, adversarial examples lead to incorrect classification, and further lead to disparate impact on the similarity between individuals as well as disparate demographic information on groups. Thus, adversarial examples prevent us from correctly assessing model fairness.

Figure 7.1 presents an example of a classification task, in which a bank makes a loan decision. The dataset consists of two groups: group 1 (triangle points) and group 0 (circle points). The fairness is evaluated in terms of individual fairness and group fairness (using demographic parity (DP) to measure the difference in positive predictions between groups [15]). Before the attack, the model satisfies individual fairness ("similar examples should be treated similarly"), and the model is evaluated with no discrimination, DP = 0. However, after the attacker converts two examples into adversarial examples by altering the number of pets owned by customers (blue triangles become black triangles), the model exhibits individual bias and group bias. Note that the number of pets is considered as a less important feature and could be imperceptible by bank experts. This shows that adversarial attacks can fool model's predictions on input examples by injecting a small amount of noise, which further skews model fairness.

In this chapter, we focus on the vulnerability of model fairness to adversarial attacks,

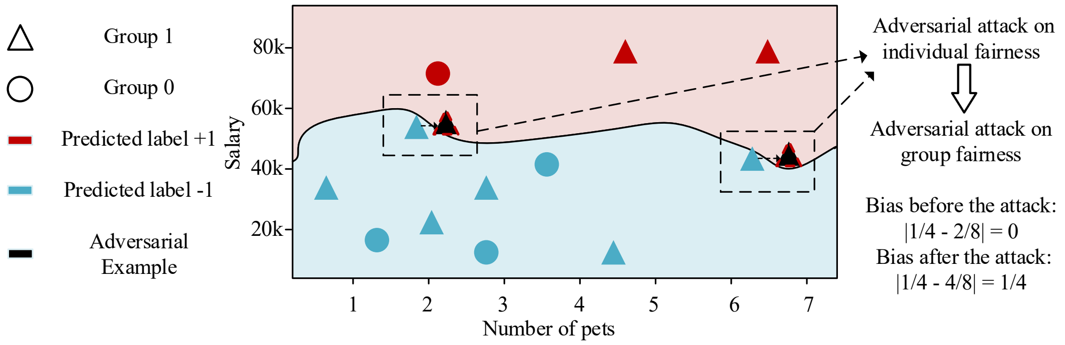

Group 1
Group 0
Predicted label +1
Predicted label -1
Adversarial Example
Salary
80k
60k
40k
20k
1
2
3
4
5
6
7
Number of pets
Adversarial attack on individual fairness
Adversarial attack on group fairness
Bias before the attack: |1/4 - 2/8| = 0
Bias after the attack: |1/4 - 4/8| = 1/4

Figure 7.1: Two examples in the group 1 with negative predicted labels are perturbed into adversarial examples. These original examples are similar to their adversarial examples but are being treated differently by the model. This violates individual fairness, and furthers skews the demographic information in group fairness.

and how to maximize the model bias with the constraint of adversarial examples and the perturbation scale. Our attacks effectively exploit the fact that data examples exhibit disparate susceptibility to fairness metrics. This can disproportionately alter the influence of data samples in different groups, which enables the attacker to place adversarial examples where it can impose a large bias on the model. Our contributions are summarized as follows:

- 1) **Vulnerability of Model Fairness:** We are the first to study the vulnerability of model fairness from the view of adversarial examples. We define individual adversarial bias and group adversarial bias, and explain the vulnerability of individual fairness and group fairness to adversarial attacks.

- 2) **Adversarial Fairness Attacks:** We formulate the adversarial fairness attack as a general optimization problem. The proposed algorithm can maximize model bias with a fixed perturbation scale on limited adversarial examples, while corporating with existing adversarial attack methods,

- 3) **Empirical Benefits:** We empirically evaluate our proposed methods across four real datasets. Results show that a small perturbation scale could skew the discrimination level dramatically.

7.2 Related Work

Adversarial Attacks Since adversarial examples were first discovered [128], it has been a hot topic in adversarial machine learning. Many successful adversarial attack methods have been proposed to generate adversarial examples, including Fast Sign Gradient Method (FSGM) [58], Deepfool [102], Projected Gradient Descend (PGD) [95], etc. Most works on adversarial attacks focus on image recognition tasks. Recently, however, adversarial attacks on tabular data have received attention [10, 18, 62, 86]. The main difficulty is the imperceptibility of perturbations in the tabular domain as opposed to the image domain. For example, Ballet et al. studied adversarial attacks on tabular data [10], and used feature importance as an indicator and applied more perturbations to the less important features. In [62], adversarial examples on tabular data are treated as counterfactual examples to help the explainability of the model. Most existing adversarial attacks aim at model accuracy. In this work, we are the first to analyze the vulnerability of model fairness via the view of adversarial attacks, and design adversarial attacks to skew model fairness.

Fair Machine Learning In recent years, many fairness metrics have been proposed to define what is fair in machine learning, including statistical fairness [31, 60, 146], individual fairness [45, 46, 117] and casual fairness [83, 137]. In this work, we design adversarial attacks aiming at individual and statistical fairness. Meanwhile, methods to achieve fair machine learning fall into three categories based on when fairness is addressed. Pre-processing methods eliminate the discrimination by transforming the training data into a fair set [16, 73, 117]. In-processing methods avoid discrimination with fair constraints or regularizers [4, 57, 145, 149]. Post-processing methods adjust the decisions made by a model to be non-discriminatory [60, 79].

Recent studies [21, 101, 103] have demonstrated that fairness is not robust to attacks. Mehrabi et al. [101] and Chang et al. [21] studied data poisoning attacks against group-based fair machine learning. Nanda et al. proved that it may be easier for an attacker to target a particular group, resulting in a form of robustness bias [103]. As can be seen, a few works have discussed fairness robustness and designed data poison attacks on group fairness. Different from these studies, we study the vulnerability of model fairness via the view of adversarial attacks. Individual fairness is naturally susceptible to adversarial examples, and further adversarial examples skew the demographic information in groups.

7.3 Vulnerability of Model Fairness to Adversarial Attacks

Notation Consider a binary classifier $f : X \to Y$, which maps training examples x to a discrete label $y \in \{0,1\}$. Let D be a dataset with N data points $D = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \sim X$ contains the information of k unprotected attributes, the protected attribute z (also called sensitive attribute), and $y_i \in \{0,1\} \sim Y$ denotes the label. When considering a binary protected attribute, D is divided into the advantage group $D_1 = \{(x_i, y_i) : z_i = 1\}_{i=1}^{N}$ and disadvantage group $D_0 = \{(x_i, y_i) : z_i = 0\}_{i=0}^{N}$. The advantage group is referred to as the group that has better performance in terms of a fairness metric and vice versa. Similarly, we have $D_1^+ = \{(x_i, y_i) : z_i = 1, y_i = 1\}_{i=1}^{N}$ and $D_0^- = \{(x_i, y_i) : z_i = 0, y_i = 0\}_{i=1}^{N}$. In particular, x_i^+ and x_i^- denote the example x_i with positive and negative labels; $\hat{x}_i^+$ and $\hat{x}_i^-$ denote the example x_i with positive and negative predicted labels.

7.3.1 Vulnerability of Individual Fairness

Individual fairness Individual fairness is formalized by viewing machine learning models as maps between input and output metric spaces and defining individual fairness as Lipschitz continuity of machine learning models. The definition is given as follows.

Definition 7.1. (**Individual Fairness**) [45] A model $f : X \to Y$ is individually fair if for all $x_i, x_j \in X$, we have $M(f(x_i), f(x_j)) \le d(x_i, x_j)$.

Individual fair models require that any two data examples x_i, x_j that are at distance $d(x_i, x_j)$, and map to distributions $f(x_i)$ and $f(x_j)$, respectively, such that the statistical distance between $f(x_i)$ and $f(x_j)$ is $d(x_i, x_j)$ at most. The distance metric d on the input space depends on the machine learning tasks because it encodes the intuition as to which users are similar.

Adversarial Attack Meanwhile, machine learning models are vulnerable to adversarial attacks. The goal of adversarial attacks is to generate an adversarial example

x' to mislead the model [128]. Given an input example (x, y), an adversarial example is a perturbation of the input pattern $x' = x + r$; here, r is small enough so that x' is imperceptible from x, but the model f predicts an incorrect label. Given the model f and L_p-norm, the adversarial example is defined by solving the following constrained optimization problem

$$x' = \underset{x':||x'-x||_p \leq \epsilon}{\operatorname{argmax}} \mathscr{L}(x', y), \tag{7.1}$$

where ϵ is the perturbation scale; $\mathscr{L}$ denotes the loss function; p denotes L_p norm. Research [34, 65] showed that the robustness of models to adversarial examples relates to the Lipschitz constant, and it can be used to establish a strict worst-case bound for model robustness.

Connection We see that Lipschitz continuity connects with individual fairness and adversarial attacks. When the perturbation norm and the similarity between examples are comparable, adversarial examples could act as similar examples for an individual. In practice, a learning model f usually does not satisfy Lipschitz condition, that is, f does not satisfy individual fairness and is not robust to adversarial attack. Adversarial examples could be the factor that leads to the break in individual fairness. John et al. have verified individual fairness in machine learning models, and found that two similar data examples could exhibit individual bias [70]. That is, a pair of valid inputs which are close to each other (according to a distance metric) but are treated differently by the model. However, it is difficult to find similar data examples in a dataset, while adversarial attack can easily produce similar data examples. To measure the impact of adversarial examples on individual fairness, we define individual adversarial bias as follows.

Definition 7.2. (Individual Adversarial Bias (IAB)) Let d be a metric to evaluate the similarity between two examples as well as the perturbation norm on adversarial examples. A model $f : X \to Y$ exhibits individual adversarial bias if there exists a pair of original and adversarial examples (x_i, x'_i), with $M(f(x_i), f(x'_i)) \geq d(x_i, x'_i)$. Such example x'_i is referred to as an individual biased example. Given a set of examples D, individual adversarial bias of model f on D is defined as

$$\Gamma_I(D, f) = Pr_{x,y \in D}\{M(f(x), f(x')) \geq d(x, x')\} \tag{7.2}$$

In the context of adversarial attacks, the constraint is used to limit the perturbation norm $d(x,x') \leq \epsilon$. Let us consider the example illustrated in Figure 1: a model f is used to make a decision as to whether an application will be accepted based on customer-provided information (incomes, age, etc.). When L_2 norm is adopted, the adversarial example x'_i can be generated by altering the number of pets, which is an unimportant feature in expert's eyes, leading to $M(f(x_i), f(x'_i)) \geq ||x_i - x'_i||_2$.

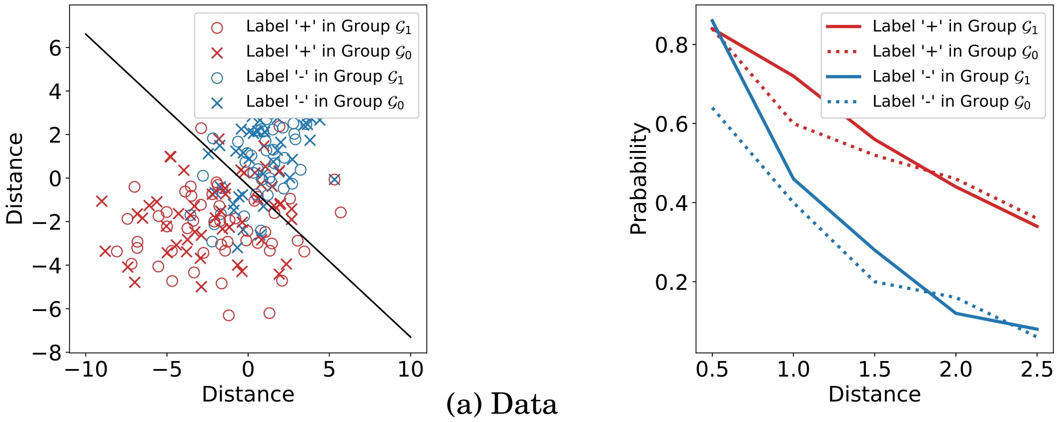
7.3.2 Vulnerability of Group Fairness

Adversarial attacks directly undermine individual fairness by generating adversarial examples. Adversarial attack can also undermine group fairness. The reasons are as follows: 1) Adversarial examples can distort the group proportion in any group fairness measure by carefully selecting adversarial examples; 2) Data examples of each group have different distances to the decision boundary. Let us take DP as an example of group fairness metrics to demonstrate the vulnerability of group fairness to adversarial attacks.

Definition 7.3. (Demographic Parity (DP)) [15] A classification model satisfies demographic parity if $Pr(f(x) = 1|z = 1) = Pr(f(x) = 1|z = 0)$, where $f(x)$ is the predicted label. The discrimination level of DP is calculated as $\mathscr{I}(D,f) = |Pr(f(x) = 1|z = 1) - Pr(f(x) = 1|z = 0)|$.

The discrimination level of DP is measured by the difference in positive predictions between groups. Intuitively, turning all examples into adversarial examples (flipping their predicted labels) will not alter the discrimination level much. This is because the full reversal of positive and negative predictions does not change the difference in the absolute value of the discrimination level.

In addition, the perturbation scale is crucial to adversarial attacks. Given a small ϵ, not all examples could be transformed into adversarial examples, and the proportion of adversarial examples in each group may be different. This is because data examples of each group exhibit disparate average distances to the decision boundary, examples in one group may be more vulnerable to adversarial attacks than those in the other group. Here, we define group adversarial bias based on DP, and quantify the disparate



(a) Data distribution of a binary classification problem for randomly generated data.

(b) Proportion of a group which is greater than ϵ away from a decision boundary

Figure 7.2: An example of group adversarial bias.

susceptibility of data examples in each group to adversarial attack with the following function.

Definition 7.4. (Group Adversarial Bias (GAB)) Given a dataset D and a perturbation scale ϵ, group adversarial bias $\Gamma_G(D,f)$ is defined as the sum of the probability that x^- in the advantage group D_1 and x^+ in the disadvantage group D_0 are perturbed into adversarial examples.

$$\Gamma_G(D,f) = Pr_{x,y\in D}\{dis(x,\mathscr{B}) \leq \epsilon | x \in D_1, y=0\} + Pr_{x,y\in D}\{dis(x,\mathscr{B}) \leq \epsilon | x \in D_0, y=1\} \tag{7.3}$$

where $dis(x,\mathscr{B})$ denotes the distance between x and the decision boundary $\mathscr{B}$. Here, we consider x^- in D_1 and x^+ in D_0 because these examples exhibit the discrimination in terms of DP. Note that $\Gamma_G(f) \in [0,1]$, and a large Γ_G indicates that the model f has the potential to be more biased by adversarial attacks.

Let us consider an example of logistic regression (LR) to show group adversarial bias. In LR, the decision boundary is well understood and can be expressed in a closed form. As a result, we can easily calculate the proximity of each point to the decision boundary. Consider a toy dataset and learned classifier in Figure 7.2 (a); here, red and blue points denote x^+ and x^-, and circle and cross denote examples in D_1 and D_0. We can quantify the disparate susceptibility by plotting the proportion of examples that are greater than ϵ from the decision boundary. As shown in Figure 7.2 (b), we observe that x^- in D_1 is generally far from the decision boundary than x^+ in D_0.

From a strictly view of classification accuracy, which class being closer to the decision boundary is not of concern because two classes could achieve similar classification accuracy. However, when we move from this toy problem to a neural network with real data, this difference between classes can become a potential vulnerability to group fairness, especially for a well-designed perturbation scale.

7.4 Method

7.4.1 Problem Formulation

Consider a white-box attacker who can access the model and dataset D. The attacker aims to maximize model bias Γ by carefully selecting a subset of examples D_s from D and converting them into a set of adversarial examples D_s' with the constraint of perturbation norm on adversarial examples. We assume that sensitive attributes are not perturbed by attackers. Because sensitive attributes are very sensitive, changes in sensitive attributes are easily discovered by domain experts. The optimization problem can be described as follows

$$\max_{D_s'} \Gamma(D', f) \tag{7.4}$$

$$s.t. \quad |D_s| \leq \eta |D| \tag{7.5}$$

$$where \quad x' = \underset{x' : ||x'-x||_p \leq \epsilon}{\operatorname{argmax}} \ L(x', y). \tag{7.6}$$

where $D' = \{D_s' \cup D_{\bar{s}}\}$ denotes the set that consists of adversarial examples and original examples; $|D_s|$ is the number of data examples that are perturbed; η denotes the proportion of data examples. The constraint (7.5) restricts the number of data examples being perturbed, and the constraint (7.6) limits the maximum perturbation on the L_p norm.

7.4.2 Adversarial Attack on Individual Fairness

The optimization goal of adversarial attacks on individual fairness is to maximize individual adversarial bias $\Gamma_I(D', f) = Pr_{x,y \in D'}\{D(f(x), f(x')) \geq d(x, x')\}$. As we can see, it depends on the metrics M and d. A description of these metrics is provided below.

Distance Metric Here, we use two distance metrics: L_2 norm and weighted L_2 norm. L_2 norm is widely used to evaluate the distance between individuals. Moreover, we use weighted L_2 norm because the imperceptibility on tabular data is relevant to feature importance [10]. Each feature x_k is associated with a feature importance coefficient $v_k \in \mathbb{R}$, and $v = \{v_1, ..., v_k\}$, and the feature importance depends on the domain knowledge of the dataset. In this way, we define weighted L_2-norm based on the product of feature importance and data in L_2 norm as $d(x_i, x'_i) = ||(x_i - x'_i) \odot v||_2$, where $\odot$ is the Hadamard product.

The choice of M depends on the form of the output. A model is fair if, for any pair of close inputs (x_i, x'_i), the model outputs are also close. In classification models, this means that the model's decision does not change (i.e., $f(x_i) = f(x'_i)$). It is therefore appropriate to use the discrete metric $M(f(x_i), f(x'_i)) = \mathbb{I}[f(x_i), f(x'_i)]$ with the threshold equal to 0 on the model output; since, in a fair classification model, we want to prevent any change in the class label due to small perturbations. The distance metric adopted in this work is a common means of measuring similarity between examples. We believe that constructing a good distance metric to measure individual adversarial bias is an important research question in the future.

Distance to Decision Boundary Given the constraints on the perturbation scale and the number of adversary examples, maximizing individual fairness bias requires the examples that are most vulnerable to adversarial attacks. Surely, the most vulnerable examples can be found by brute-force search. However, as we know, the generation of adversarial examples is time-consuming. As the community has discovered, data examples closer to the decision boundary might be more susceptible to adversarial attack [64, 129, 134]. Some studies have conducted empirical research to show this distance, but there is no explicit way to measure it. We use an approximate method to calculate the distance between examples and the decision boundary. Let $f_0(x)$ and $f_1(x)$ be the outputs of the softmax layer of a neural network. The decision boundary of the network can be defined as $\mathscr{B} = \{x | f_0(x) = f_1(x)\}$. According to [50], the distance to the boundary can be approximated as

$$dis(x, \mathscr{B}) = \frac{|f_0(x) - f_1(x)|}{\|\nabla_x f_0(x) - \nabla_x f_1(x)\|_p} \tag{7.7}$$

Eq.(7) can be used to calculate the approximate distance between examples and the decision boundary. The optimization problem of individual adversarial bias is then solved by selecting the examples closest to the decision boundary. When an example is near the decision boundary, the distance from the example x can be further approximated as $dis(x,\mathscr{B}) = |g_0(x) - g_1(x)|$ according to [120] where $g_0(x)$ and $g_1(x)$ denote the inputs to the softmax layer for classes 0 and 1. Thus, given a perturbation scale ϵ, if $|g_0(x) - g_1(x)| \leq \epsilon$, we can consider this sample has the potential to produce adversarial examples that undermine individual fairness.

7.4.3 Adversarial Attack on Group Fairness

Intuitively, the goal of maximizing group adversarial bias is to turn $\hat{x}^-$ into $\hat{x}^+$ in G_1, and turn $\hat{x}^+$ into $\hat{x}^-$ in G_0 to the greatest extent. In this way, the distinction in positive predictions between groups increases, leading to a more biased model.

Objective Function We now transform group adversarial bias Γ_G defined in Eq. (7.3) according to the definition of DP, which can be expressed as

$$\begin{aligned}(7.8)\qquad &\Gamma_G(D', f)\\ &= Pr_{x,y\in D'}\{dis(x,\mathscr{B})) \leq \epsilon | x \in D_1, y = 0\} + Pr_{x,y\in D'}\{dis(x,\mathscr{B})) \leq \epsilon | x \in D_0, y = 1\}\\ &= \frac{\sum_{i=1}^{|D'|} Pr[f(x_i) = 1](1 - y_i)}{|D_1|} + \frac{\sum_{i=1 D'|}(1 - Pr[f(x_i) = 1])y_i}{|D_0|}\end{aligned}$$

where $|D_1|$ and $|D_0|$ denote the number of examples in D_1 and D_0. Note that $|D_1| = |D'_1|$ and $|D_0| = |D'_0|$ because the sensitive attribute of each example does not change. Next, we propose an algorithm to solve the optimization problem.

Adversarial Example Selection Given the perturbation scale, ϵ and the number of perturbed examples $|D_s|$, the value of group adversarial bias depends on the likelihood of an example being converted into an adversarial example according to Eq.(7.8). To maximize group adversarial bias with constraints, we need to select examples that are more likely to become adversarial examples. In other words, we need to select examples in D_1^- and D_0^+ that are closest to the decision boundary. Note that group adversarial bias is the same when one $\hat{x}^+$ is perturbed into $\hat{x}^-$ in D_0 and one $\hat{x}^-$ is perturbed into

$\hat{x}^+$ in group D_1. First, we gather examples in D_1^- and D_0^+, and rank them in terms of the distance $dis(x,\mathscr{B})$ calculated in Eq. (7.7) in an ascending order. We can then select data examples according to the rule in Eq. (7.9). Algorithm 1 describes the process of the optimization problem of maximizing group adversarial bias.

$$D_s = \begin{cases} \eta|D|, & if \quad |D_s| \leq |D_1^-| + |D_0^+| \\ |D_1^-| + |D_0^+|, & Otherwise \end{cases} \tag{7.9}$$

Theorem 7.1. *Let $\mathscr{I}(D,f)$ and $\mathscr{I}(D',f)$ denote the discrimination level of model f on the dataset D and D'. The range of $\mathscr{I}(D',f)$ with our proposed method is as follows*

$$0 \leq \mathscr{I}(D,f) \leq \mathscr{I}(D',f) \leq \mathscr{I}(D,f) + \frac{|D_{s,1}^-|}{|D_1|} + \frac{|D_{s,0}^+|}{|D_0|} \leq 1, \tag{7.10}$$

where $|D_{s,1}^-|$ and $|D_{s,0}^+|$ denote the number of examples in $D_{s,1}^-$ and $D_{s,0}^+$. Theorem 7.1 states that the selection of adversarial examples with our method will increase the discrimination level, and the proof is provided in the Supplementary.

Proof. First, we decompose the discrimination level of $\mathscr{I}(D',f)$ in terms of the definition of DP.

(7.11)

$$\begin{aligned}
&\mathscr{I}(D',f) \\
&= |Pr(f(x)=1|z=1) - Pr(f(x)=1|z=0)| \\
&= Pr(f(x)=1|z=1) - Pr(f(x)=1|z=0) \\
&= \frac{\sum_{i=1}^{|D|} f(x_i)z_i + \sum_{i=1}^{|D'_s|} Pr[f(x'_i)=1]z_i}{|D_1|} - \frac{\sum_{i=1}^{|D|}(1-f(x_i))(1-z_i) - \sum_{i=1}^{|D'_s|}(1-Pr[f(x'_i)=1])(1-z_i))}{|D_0|} \\
&= (\frac{\sum_{i=1}^{|D|} f(x_i)z_i}{|D_1|} - \frac{\sum_{i=1}^{|D|}(1-f(x_i))(1-z_i)}{|D_0|}) + (\frac{\sum_{i=1}^{|D'_s|} Pr[f(x'_i)=1]z_i}{|D_1|} + \frac{\sum_{i=1}^{|D'_s|}(1-Pr[f(x'_i)=1])(1-z_i)}{|D_0|}) \\
&\leq \mathscr{I}(D,f) + \frac{|D_{s,1}^-|}{|D_1|} + \frac{|D_{s,0}^+|}{|D_0|}
\end{aligned}$$

The absolute value of $\mathscr{I}(D',f)$ is removed because when the adversary has the background knowledge of which group is the advantaged group. Positive predictions in the advantage group come from two parts: the original set D and the adversarial set

D'_s. While positive predictions in the disadvantaged group come from the original set D minus the adversarial set D'_s. The minus is because some examples with positive predictions are turned into adversarial examples with negative predictions. Note that $|D_1| = |D'_1|$ and $|D_0| = |D'_0|$ because the sensitive attribute does not change. The last inequation is because not all selected examples can flip predicted labels successfully. Thus the upper bound of discrimination level is when all selected examples in $D^+_{s,0}$ and $D^-_{s,1}$ flip their predicted labels. The lower bound of $\mathscr{I}(D',f)$ is $\mathscr{I}(D,f)$, which means that selected examples fail to flip predicted labels given some hard conditions, such as a small perturbation scale. ■

Algorithm 10 Adversarial Example Selection

```
Input: Dataset $D$, the number of adversarial examples $|D'_s|$, perturbation scale $\epsilon$
  for $x_i \in D$ do
    if $x_i \in D^-_1$ then
      Execute an adversarial attack method to generate $x'_i$
      Calculate the distance $dis(x, \mathscr{B})$ according to Eq.(7.7)
    end if
    if $x_i \in D^+_0$ then
      Execute an adversarial attack method to generate $x'_i$
      Calculate the distance $dis(x, \mathscr{B})$ according to Eq.(7.7)
    end if
  end for
  Sort $x'$ according to distance $dis(x, \mathscr{B})$ in an increasing order
  Select $x'$ according to Eq. (7.9)
Output: A set of adversarial examples $D'_s$
```

7.5 Experiment

The aim of our experiments is to assess: the vulnerability of model fairness to adversarial attacks, and the effectiveness of our methods to individual and group fairness in terms of the discrimination level and the perturbation norm.

Table 7.1: Description of datasets

Dataset	Size	No. features	Sensitive attribute	Prediction task	Discrimination
German	1000	20	Age	Approve loan application?	0.0407
Bank	45211	20	Age	Subscribed to a deposit account?	0.0627
Titanic	891	9	Race	Survive?	0.0081
Law	1823	17	Gender	Pass the exam?	0.0572

7.5.1 Experimental Setup

7.5.1.1 Datasets, Models and Parameters

We demonstrate our proposed methods on four datasets well used in the fairness literature: German, Bank, Titanic and Law. The basic description of these datasets is given in Table 7.1. All datasets are preprocessed using uniform sampling [73] to obtain fair datasets in terms of demographic parity. As shown in Table 7.1, the discrimination level in four datasets is small.

We build a fully connected neural network using dense ReLU layers and a Softmax layer for the last one. We split the dataset into training set (70%), and test set (30%). We set the size of selected examples $D_s = 0.2 * D$; the perturbation scale $\epsilon = 0.4$ on the Titanic dataset and $\epsilon = 0.2$ on the other three datasets.We use the Adam optimizer with learning rate 0.001, batch size 64, momentum 0.9 and weight decay 0.001 to train neural networks. To limit the generated adversarial examples x' in a coherent subspace, we clip x' to the bounds of each feature. More formally, we clip each feature $x_i \in [1...k]$ so that it remains within its their natural range $[min(x_i), max(x_i)]$.

7.5.1.2 Evaluation Metrics

Adversarial Bias: Individual adversarial bias (IAB) is measured by the success rate of individual biased examples by Eq. (7.2), and group adversarial bias (GAB) is measured by Eq. (7.3).

Perturbation Norms: To evaluate the degree to which the adversarial attack is imperceptible, we measure the non-weighted norm of the adversarial perturbation $||x - x'||_2$ and weighted norm of the adversarial perturbation $||(x - x') \odot v||_2$. In both perturbation norms, we compute the mean and median value over the set of adversarial examples D'_s. We denote the mean and median of the perturbation norm as $MeanP$

and $WMeanP$; the weighted mean and median of the weighed perturbation norm as $WMeanP$ and $WMedian$.

Distance to Closest Neighbor: Although the perturbation norm of an example allows us to compare the scale of perturbation with itself, this value does not provide insight into the importance of perturbation for a given dataset. We, therefore, compute the average non-weighted and weighted distance between the adversarial example and the closest neighbors from the original samples. Thus, we can measure the effect of the perturbation on the dataset. The distance to the closest neighbor is defined as $d^n(x_i') = \underset{x \in D}{\operatorname{argmin}} ||x_i' - x||_2$ and $d^{wn}(x_i') = \underset{x \in D}{\operatorname{argmin}} ||(x_i' - x) \odot v||_2$ where x denotes the original examples in the dataset D. We use MD and WMD to denote the non-weighted and weighted mean distance between an adversarial example to its closest neighbors. In the experiment, we calculate the average distance of the nearest three neighbors.

7.5.1.3 Methods and Baselines

In this experiment, we consider two methods, **LowProFool (LPF)** and **DeepFool (DF)**, to generate adversarial examples on tabular data. A brief description of these methods is presented below; more details of these methods can be found in the Supplementary. **LPF** generates adversarial examples on tabular data by making use of feature importance as an indicator and applies more perturbations to the less important features [10]. The perceptibility of LPF is measured as the L_2 norm of the perturbation weighted by a feature importance vector. **DF** calculates the minimal perturbation to change the classifier‚Äôs decision according to the orthogonal projection of original examples onto the hyperparameter plane [102]. The perturbed example updates while the true label and the label of the perturbed example are the same. We use **LPF-Pro** and **DF-Pro** to denote LPF and DF implemented using our proposed methods. Considering it is the first paper to work on adversarial fairness attack, the baseline methods are implemented with randomly selected examples used as adversarial examples, which are denoted by **LPF-Ran** and **DF-Ran**.

7.5.2 Experimental Results

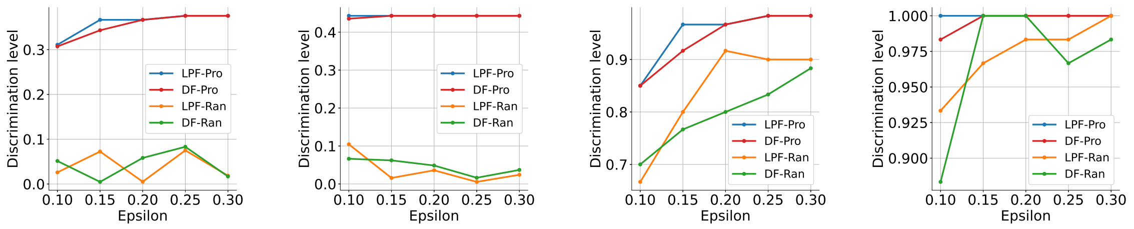
Table 7.2 presents the performance of all methods on four datasets. Our findings show that the proposed methods exhibit significantly higher discrimination and a lower

Table 7.2: Results for all datasets and all methods concerning fairness metrics and perturbation metrics

DATA SET	METHOD	IAB	GAB	*MeanP*	*WMeanP*	*MedianP*	*WMedianP*	*MD*	*WMD*
GERMAN	LPF-PRO	**0.9967**	**0.3665**	0.1805	**0.0242**	0.2039	**0.0151**	1.5161	1.6195
	DF-PRO	**0.9667**	**0.3665**	**0.1060**	0.0339	**0.0878**	0.0253	**1.3317**	**1.3464**
	LPF-RAN	0.9167	0.0051	0.2091	0.0326	0.2304	0.0207	1.4970	1.5767
	DF-RAN	0.8000	0.0582	0.1464	0.0447	0.1273	0.0474	1.5636	1.5979
BANK	LPF-PRO	**1.0000**	**0.4430**	0.1632	0.0214	0.1619	0.0172	0.7114	1.0749
	DF-PRO	**1.0000**	**0.4430**	**0.0519**	**0.0212**	**0.0410**	**0.0170**	**0.6912**	**1.0629**
	LPF-RAN	0.9833	0.0359	0.1895	0.0347	0.1760	0.0290	1.9914	1.6412
	DF-RAN	1.000	0.0485	0.0956	0.0391	0.0953	0.0392	1.0568	1.8285
TITANIC	LPF-PRO	**1.0000**	**0.3784**	0.3072	**0.2742**	0.3111	**0.2788**	**0.3280**	**0.4702**
	DF-PRO	**1.0000**	**0.3784**	**0.3071**	0.2751	**0.3105**	0.2792	0.3282	0.4705
	LPF-RAN	0.7833	0.0238	0.3159	0.2829	0.3135	0.2811	1.0099	1.0184
	DF-RAN	0.8167	0.0115	0.3142	0.2821	0.3126	0.2801	1.0918	0.0751
LAW	LPF-PRO	0.9700	0.2951	0.0662	**0.0191**	0.0495	**0.0126**	1.0427	1.1446
	DF-PRO	**0.9800**	**0.2984**	**0.0578**	0.0203	**0.0398**	0.0145	**1.0056**	**1.0931**
	LPF-RAN	0.8200	0.0861	0.0981	0.0294	0.1015	0.0263	1.1660	1.2145
	DF-RAN	0.8900	0.0562	0.0819	0.0303	0.0715	0.0251	1.0626	1.1820

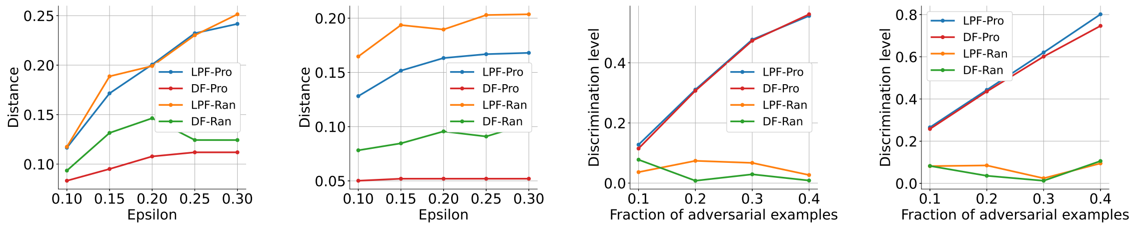
perturbation norm than the baselines. As shown, LPE-Pro and DF-Pro can generate huge individual and group adversarial bias. For examples, group adversarial bias is 0.3665 in LPF-Pro and DF-Pro with the German dataset. While group adversarial bias in LPF-Ran and DF-Ran are quite small, less than 0.1. This indicates that adversarial examples can effectively turn a fair model into an unfair model with our proposed methods. As for the perturbation norm, $MeanP$, $WMeanP$, $MedianP$ and $WMedianP$ are lower in LPF-Pro and DF-Pro than in LPF-Ran and DF-Ran. This indicates that selected examples in the proposed methods can be perturbed into adversarial examples with fewer perturbations than randomly selected examples. Also, $Wmean$ and $WMedian$ are typically larger in LPF than DF, while $WmeanP$ and $WMedianP$ are smaller in LPF than DF. This indicates that LPF performs a lower perturbation norm when considering feature importance as an indicator of imperceptibility. MD and WMD show the perturbation norm by comparing with its nearest neighbors; we set 3 nearest neighbors in the experiment. As shown, the perturbation norm is much smaller than the distance to the nearest neighbors, which further highlights that the perturbation norm is imperceptible.

Ablation Study Next, to better understand how adversarial examples affect model fairness in the proposed methods, we show the performance with varying perturbation



(a) German-GAB (b) Bank-GAB (c) German-IAB (d) Bank-IAB

Figure 7.3: Group adversarial bias and individual group adversarial bias on the German and Bank datasets with increasing perturbation scale ϵ. Results show that adversarial examples in our proposed methods can significantly skew individual fairness and group fairness more than baselines.



(a) German-MeanP (b) Bank-MeanP (c) German-GAB (d) Bank-GAB

Figure 7.4: The mean of the perturbation norm on the German and Bank datasets with an increasing number of adversarial examples ϵ in Fig. 7.4 (a) and 7.4 (b). The result of the group adversarial bias on the German and Bank datasets with the growing number of adversarial examples in Fig. 7.4 (c) and 7.4 (d).

scale ϵ and the number of adversarial examples. Figure 7.3 shows group and individual adversarial bias with varying perturbation scale ϵ on the German and Bank datasets (The results for all datasets are given in the supplementary). The discrimination level starts to increase and then becomes steady with increasing ϵ. This is because with a larger perturbation norm, examples are easier to find its adversarial example. Furthermore, group adversarial bias grows when more adversarial examples are present. However, a larger ϵ does not help LPF-Ran and DF-Ran generate more group adversarial bias.

In Figure 7.4 (a) and 7.4 (b), the mean of the perturbation norm $MeanP$ increases with a growing ϵ. Meanwhile, $MeanP$ in the proposed methods is smaller the baselines, which proves that our proposed methods select examples that are more vulnerable to adversarial attacks. Figures 7.4 (c) and 7.4 (d) show the changes in group adversarial bias with a varying number of adversarial examples. Group adversarial bias in the proposed methods increases as the number of adversarial examples increases. This is because the proposed methods generate adversarial examples in the direction where the discrimination increases. However, the discrimination level in the baselines does not change much. This is because randomly generated adversarial examples do not lead to a preference for a specific group.

7.6 Summary

In this chapter, we analyze the vulnerability of model fairness via the view of adversarial examples. Our proposed algorithms are able to maximize the discrimination level with the constraints of a number of adversarial examples and the perturbation scale. In practice, the proposed algorithms are very general and are straightforward to apply in any adversarial attacks to skew model fairness. We hope the insights revealed will drive further research into solutions for building robustly fair learning models.

CHAPTER 8

CONCLUSION AND FUTURE WORK

8.1 Conclusion

This thesis broadly investigates privacy and fairness problems in machine learning. The proposed methods and experimental results suggest that machine learning can provide users with a better experience in terms of privacy and fairness without sacrificing too much model performance. The results from Chapter 3 show that correlated datasets may leak more privacy information than expected. A careful selection of features will help achieve a better trade-off between data utility and data privacy in correlated datasets. Chapters 4 and 5 show that unlabeled data is able to improve the trade-off between fairness and accuracy in machine learning, and pre-processing and in-processing methods in supervised learning can also be applied in semi-supervised learning with some adjustments. Chapter 6 shows that DP-SGD makes the training less stable and has a disparate impact on model fairness, and the discrimination level can be used as a key indicator to guide when to stop training. We view the results as a promising preliminary investigation into the relationships between training time and model balance. Finally, Chapter 7 shows the vulnerability of model fairness via the view of adversarial examples. Our key finding is that adversarial examples easily affect individual fairness, and further adversarial examples will distort demographic information among groups.

8.2 Future Work

When working on fair semi-supervised learning, we have an assumption that labeled and unlabeled have very similar data distributions. However, this assumption may not hold in some real-world practices. When labeled and unlabeled data distributions are unknown or different, the bias estimation could be inaccurate from labeled data to unlabeled data. Hence, one of the future research directions is how to achieve fair semi-supervised learning where labeled and unlabeled data have different data distributions. Another further direction is to explore ways to achieve fair semi-supervised learning with other fairness metrics, such as individual and casual fairness. The dominant research focus in fair learning is known as group fairness. However, individual and casual fairness catch types of unfairness that group fairness metrics miss. individual and casual fairness in semi-supervised learning is an open question of how unlabeled data may also benefit other fairness metrics.

Interaction between privacy and fairness Most existing work assumes that sensitive attributes are known when building fairness-aware models and evaluating fairness metrics. For example, sensitive attributes of gender and race are known information when designing fairness constraints. However, sensitive attributes are usually considered as sensitive information which is unavailable because of privacy concerns or the dataset may be constructed for use in a situation where the sensitive information is collected unnecessarily, undesired, or even illegally. In many situations, collecting large datasets with such sensitive information is impossible. Therefore, training and evaluation of fair machine learning models become a challenging problem. This is an open question of how to build fairness-aware and evaluate fair machine learning without sensitive attributes.

BIBLIOGRAPHY

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, *Deep learning with differential privacy*, in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016, pp. 308–318.

[2] A. Abid, M. Farooqi, and J. Zou, *Persistent anti-muslim bias in large language models*, arXiv preprint arXiv:2101.05783, (2021).

[3] A. Acar, H. Aksu, A. S. Uluagac, and M. Conti, *A survey on homomorphic encryption schemes: Theory and implementation*, ACM Computing Surveys (CSUR), 51 (2018), pp. 1–35.

[4] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, *A reductions approach to fair classification*, arXiv preprint arXiv:1803.02453, (2018).

[5] S. Aghaei, M. J. Azizi, and P. Vayanos, *Learning optimal and fair decision trees for non-discriminative decision-making*, arXiv preprint arXiv:1903.10598, (2019).

[6] R. Agrawal and R. Srikant, *Privacy-preserving data mining*, in Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 2000, pp. 439–450.

[7] K. Amin, A. Kulesza, A. Munoz, and S. Vassilvtiskii, *Bounding user contributions: A bias-variance trade-off in differential privacy*, in International Conference on Machine Learning, 2019, pp. 263–271.

[8] A. BACKURS, P. INDYK, K. ONAK, B. SCHIEBER, A. VAKILIAN, AND T. WAGNER, *Scalable fair clustering*, arXiv preprint arXiv:1902.03519, (2019).

[9] E. BAGDASARYAN, O. POURSAEED, AND V. SHMATIKOV, *Differential privacy has disparate impact on model accuracy*, in Advances in Neural Information Processing Systems, 2019, pp. 15453–15462.

[10] V. BALLET, X. RENARD, J. AIGRAIN, T. LAUGEL, P. FROSSARD, AND M. DETYNIECKI, *Imperceptible adversarial attacks on tabular data*, arXiv preprint arXiv:1911.03274, (2019).

[11] S. BERA, D. CHAKRABARTY, N. FLORES, AND M. NEGAHBANI, *Fair algorithms for clustering*, in Advances in Neural Information Processing Systems 32, 2019, pp. 4955–4966.

[12] S. L. BLODGETT, L. GREEN, AND B. O,ÄôCONNOR, *Demographic dialectal variation in social media: A case study of african-american english*, in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 1119–1130.

[13] S. L. BLODGETT, J. WEI, AND B. O,ÄôCONNOR, *Twitter universal dependency parsing for african-american and mainstream american english*, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 1415–1425.

[14] L. BREIMAN, *Bagging predictors*, Machine learning, 24 (1996), pp. 123–140.

[15] T. CALDERS, F. KAMIRAN, AND M. PECHENIZKIY, *Building classifiers with independency constraints*, in 2009 IEEE International Conference on Data Mining Workshops, IEEE, 2009, pp. 13–18.

[16] F. CALMON, D. WEI, B. VINZAMURI, K. N. RAMAMURTHY, AND K. R. VARSHNEY, *Optimized pre-processing for discrimination prevention*, in Advances in Neural Information Processing Systems, 2017, pp. 3992–4001.

[17] ——, *Optimized pre-processing for discrimination prevention*, in Advances in Neural Information Processing Systems, 2017, pp. 3992–4001.

[18] F. CARTELLA, O. ANUNCIACAO, Y. FUNABIKI, D. YAMAGUCHI, T. AKISHITA, AND O. ELSHOCHT, *Adversarial attacks for tabular data: Application to fraud detection and imbalanced data*, arXiv preprint arXiv:2101.08030, (2021).

[19] A. CASTELNOVO, R. CRUPI, G. GRECO, AND D. REGOLI, *The zoo of fairness metrics in machine learning*, arXiv preprint arXiv:2106.00467, (2021).

[20] G. CHANDRASHEKAR AND F. SAHIN, *A survey on feature selection methods*, Computers & Electrical Engineering, 40 (2014), pp. 16–28.

[21] H. CHANG, T. D. NGUYEN, S. K. MURAKONDA, E. KAZEMI, AND R. SHOKRI, *On adversarial bias and the robustness of fair machine learning*, arXiv preprint arXiv:2006.08669, (2020).

[22] H. CHANG AND R. SHOKRI, *On the privacy risks of algorithmic fairness*, arXiv preprint arXiv:2011.03731, (2020).

[23] K. CHAUDHURI AND C. MONTELEONI, *Privacy-preserving logistic regression.*, in NIPS, vol. 8, Citeseer, 2008, pp. 289–296.

[24] ——, *Privacy-preserving logistic regression*, in Advances in neural information processing systems, 2009, pp. 289–296.

[25] K. CHAUDHURI, C. MONTELEONI, AND A. D. SARWATE, *Differentially private empirical risk minimization*, Journal of Machine Learning Research, 12 (2011), pp. 1069–1109.

[26] I. CHEN, F. D. JOHANSSON, AND D. SONTAG, *Why is my classifier discriminatory?*, in Advances in Neural Information Processing Systems 31, 2018, pp. 3539–3550.

[27] R. CHEN, B. C. FUNG, S. Y. PHILIP, AND B. C. DESAI, *Correlated network data publication via differential privacy*, The VLDB Journal, 23 (2014), pp. 653–676.

[28] X. CHEN, B. FAIN, C. LYU, AND K. MUNAGALA, *Proportionally fair clustering*, in ICML, 2019.

[29] X. CHEN, B. FAIN, L. LYU, AND K. MUNAGALA, *Proportionally fair clustering*, in International Conference on Machine Learning, PMLR, 2019, pp. 1032–1041.

[30] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii, *Fair clustering through fairlets*, in Advances in Neural Information Processing Systems, 2017, pp. 5029–5037.

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

[32] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil, *Leveraging labeled and unlabeled data for consistent fair binary classification*, in Advances in Neural Information Processing Systems, 2019, pp. 12739–12750.

[33] ——, *Leveraging labeled and unlabeled data for consistent fair binary classification*, in Advances in Neural Information Processing Systems, 2019, pp. 12739–12750.

[34] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier, *Parseval networks: Improving robustness to adversarial examples*, in Proceedings of the 34th International Conference on Machine Learning, D. Precup and Y. W. Teh, eds., vol. 70 of Proceedings of Machine Learning Research, PMLR, 06–11 Aug 2017, pp. 854–863.

[35] M. Coşkun, A. Uçar, Ö. Yildirim, and Y. Demir, *Face recognition based on convolutional neural network*, in 2017 International Conference on Modern Electrical and Energy Systems (MEES), IEEE, 2017, pp. 376–379.

[36] A. Cotter, H. Jiang, and K. Sridharan, *Two-player games for efficient non-convex constrained optimization*, arXiv preprint arXiv:1804.06500, (2018).

[37] E. Creager, D. Madras, T. Pitassi, and R. Zemel, *Causal modeling for fairness in dynamical systems*, in International Conference on Machine Learning, PMLR, 2020, pp. 2185–2195.

[38] R. Cummings, V. Gupta, D. Kimpara, and J. Morgenstern, *On the compatibility of privacy and fairness*, in Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, 2019, pp. 309–315.

[39] A. DA,ÄÔU AND N. SALIM, *Recommendation system based on deep learning methods: a systematic review and new directions*, Artificial Intelligence Review, 53 (2020), pp. 2709–2748.

[40] J. DING, X. ZHANG, X. LI, J. WANG, R. YU, AND M. PAN, *Differentially private and fair classification via calibrated functional mechanism*, arXiv preprint arXiv:2001.04958, (2020).

[41] M. F. DIXON, I. HALPERIN, AND P. BILOKON, *Machine Learning in Finance*, Springer, 2020.

[42] P. DOMINGOS, *A unified bias-variance decomposition*, in Proceedings of 17th International Conference on Machine Learning, 2000, pp. 231–238.

[43] M. DONINI, L. ONETO, S. BEN-DAVID, J. S. SHAWE-TAYLOR, AND M. PONTIL, *Empirical risk minimization under fairness constraints*, in Advances in Neural Information Processing Systems, 2018, pp. 2791–2801.

[44] C. DWORK, *Differential privacy: A survey of results*, International conference on theory and applications of models of computation, (2008), pp. 1–19.

[45] C. DWORK, M. HARDT, T. PITASSI, O. REINGOLD, AND R. ZEMEL, *Fairness through awareness*, in Proceedings of the 3rd innovations in theoretical computer science conference, ACM, 2012, pp. 214–226.

[46] C. DWORK, C. ILVENTO, AND M. JAGADEESAN, *Individual fairness in pipelines*, arXiv preprint arXiv:2004.05167, (2020).

[47] C. DWORK, K. KENTHAPADI, F. MCSHERRY, I. MIRONOV, AND M. NAOR, *Our data, ourselves: Privacy via distributed noise generation*, in Advances in Cryptology - EUROCRYPT, 2006, pp. 486–503.

[48] C. DWORK AND A. ROTH, *The algorithmic foundations of differential privacy*, Foundations and Trends in Theoretical Computer Science, 9 (2014), pp. 211–407.

[49] H. EDWARDS AND A. STORKEY, *Censoring representations with an adversary*, arXiv preprint arXiv:1511.05897, (2015).

[50] G. F. Elsayed, D. Krishnan, H. Mobahi, K. Regan, and S. Bengio, *Large margin deep networks for classification*, arXiv preprint arXiv:1803.05598, (2018).

[51] Z. Erkin, J. R. Troncoso-Pastoriza, R. L. Lagendijk, and F. Pérez-González, *Privacy-preserving data aggregation in smart metering systems: An overview*, IEEE Signal Processing Magazine, 30 (2013), pp. 75–86.

[52] R. Feng, Y. Yang, Y. Lyu, C. Tan, Y. Sun, and C. Wang, *Learning fair representations via an adversarial framework*, arXiv preprint arXiv:1904.13341, (2019).

[53] M. Fredrikson, S. Jha, and T. Ristenpart, *Model inversion attacks that exploit confidence information and basic countermeasures*, in Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015, pp. 1322–1333.

[54] K. Fukuchi, Q. K. Tran, and J. Sakuma, *Differentially private empirical risk minimization with input perturbation*, in International Conference on Discovery Science, Springer, 2017, pp. 82–90.

[55] B. C. Fung, K. Wang, R. Chen, and P. S. Yu, *Privacy-preserving data publishing: A survey of recent developments*, ACM Computing Surveys (Csur), 42 (2010), pp. 1–53.

[56] J. Geumlek, S. Song, and K. Chaudhuri, *Renyi differential privacy mechanisms for posterior sampling*, in Advances in Neural Information Processing Systems, 2017, pp. 5289–5298.

[57] G. Goh, A. Cotter, M. Gupta, and M. P. Friedlander, *Satisfying real-world goals with dataset constraints*, in Advances in Neural Information Processing Systems 29, 2016, pp. 2415–2423.

[58] I. Goodfellow, J. Shlens, and C. Szegedy, *Explaining and harnessing adversarial examples*, in International Conference on Learning Representations, 2015.

[59] P. Gordaliza, E. Del Barrio, G. Fabrice, and J.-M. Loubes, *Obtaining fairness using optimal transport theory*, in International Conference on Machine Learning, PMLR, 2019, pp. 2357–2365.

[60] M. Hardt, E. Price, N. Srebro, et al., *Equality of opportunity in supervised learning*, in Advances in neural information processing systems, 2016, pp. 3315–3323.

[61] I. A. T. Hashem, V. Chang, N. B. Anuar, K. Adewole, I. Yaqoob, A. Gani, E. Ahmed, and H. Chiroma, *The role of big data in smart city*, International Journal of information management, 36 (2016), pp. 748–758.

[62] M. Hashemi and A. Fathi, *Permuteattack: Counterfactual explanation of machine learning credit scorecards*, arXiv preprint arXiv:2008.10138, (2020).

[63] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[64] W. He, B. Li, and D. Song, *Decision boundary analysis of adversarial examples*, in International Conference on Learning Representations, 2018.

[65] M. Hein and M. Andriushchenko, *Formal guarantees on the robustness of a classifier against adversarial manipulation*, in Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., vol. 30, Curran Associates, Inc., 2017.

[66] L. Huang and N. Vishnoi, *Stable and fair classification*, in International Conference on Machine Learning, PMLR, 2019, pp. 2879–2890.

[67] M. Jagielski, M. Kearns, J. Mao, A. Oprea, A. Roth, S. Sharifi-Malvajerdi, and J. Ullman, *Differentially private fair learning*, arXiv preprint arXiv:1812.02696, (2018).

[68] J. Janai, F. Güney, A. Behl, A. Geiger, et al., *Computer vision for autonomous vehicles: Problems, datasets and state of the art*, Foundations and Trends® in Computer Graphics and Vision, 12 (2020), pp. 1–308.

[69] D. Ji, P. Smyth, and M. Steyvers, *Can i trust my fairness metric? assessing fairness with unlabeled data and bayesian inference*, in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds., vol. 33, Curran Associates, Inc., 2020, pp. 18600–18612.

[70] P. G. John, D. Vijaykeerthy, and D. Saha, *Verifying individual fairness in machine learning models*, in Conference on Uncertainty in Artificial Intelligence, PMLR, 2020, pp. 749–758.

[71] A. Joulin, É. Grave, P. Bojanowski, and T. Mikolov, *Bag of tricks for efficient text classification*, in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, 2017, pp. 427–431.

[72] C. Jung, M. Kearns, S. Neel, A. Roth, L. Stapleton, and Z. S. Wu, *Eliciting and enforcing subjective individual fairness*, arXiv preprint arXiv:1905.10660, (2019).

[73] F. Kamiran and T. Calders, Knowledge and Information Systems, 33 (2012), pp. 1–33.

[74] ——, *Data preprocessing techniques for classification without discrimination*, Knowledge and Information Systems, 33 (2012), pp. 1–33.

[75] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, *Fairness-aware classifier with prejudice remover regularizer*, in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2012, pp. 35–50.

[76] D. Kifer and A. Machanavajjhala, *No free lunch in data privacy*, in Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, 2011, pp. 193–204.

[77] ——, *Pufferfish: A framework for mathematical privacy definitions*, ACM Transactions on Database Systems (TODS), 39 (2014), p. 3.

[78] N. KILBERTUS, M. R. CARULLA, G. PARASCANDOLO, M. HARDT, D. JANZING, AND B. SCHÖLKOPF, *Avoiding discrimination through causal reasoning*, in Advances in Neural Information Processing Systems, 2017, pp. 656–666.

[79] M. P. KIM, A. GHORBANI, AND J. ZOU, *Multiaccuracy: Black-box post-processing for fairness in classification*, in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 247–254.

[80] M. KLEINDESSNER, P. AWASTHI, AND J. MORGENSTERN, *Fair k-center clustering for data summarization*, in Proceedings of the 36th International Conference on Machine Learning, vol. 97, Long Beach, California, USA, 09–15 Jun 2019, pp. 3448–3457.

[81] M. KLEINDESSNER, S. SAMADI, P. AWASTHI, AND J. MORGENSTERN, *Guarantees for spectral clustering with fairness constraints*, in Proceedings of the 36th International Conference on Machine Learning, vol. 97, Long Beach, California, USA, 09–15 Jun 2019, pp. 3458–3467.

[82] J. KOMIYAMA, A. TAKEDA, J. HONDA, AND H. SHIMAO, *Nonconvex optimization for regression with fairness constraints*, in International conference on machine learning, 2018, pp. 2737–2746.

[83] M. J. KUSNER, J. LOFTUS, C. RUSSELL, AND R. SILVA, *Counterfactual fairness*, in Advances in Neural Information Processing Systems, 2017, pp. 4066–4076.

[84] A. LAMY, Z. ZHONG, A. K. MENON, AND N. VERMA, *Noise-tolerant fair classification*, in Advances in Neural Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds., vol. 32, Curran Associates, Inc., 2019.

[85] D.-H. LEE, *Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks*, in Workshop on Challenges in Representation Learning, ICML, vol. 3, 2013, p. 2.

[86] E. LEVY, Y. MATHOV, Z. KATZIR, A. SHABTAI, AND Y. ELOVICI, *Not all datasets are born equal: On heterogeneous data and adversarial examples*, arXiv preprint arXiv:2010.03180, (2020).

[87] M. Li, M. Soltanolkotabi, and S. Oymak, *Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks*, in International Conference on Artificial Intelligence and Statistics, 2020, pp. 4313–4324.

[88] C. Liu, S. Chakraborty, and P. Mittal, *Dependence makes you vulnberable: Differential privacy under dependent tuples.*, in NDSS, vol. 16, 2016, pp. 21–24.

[89] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt, *Delayed impact of fair machine learning*, in Proceedings of the 35th International Conference on Machine Learning, vol. 80 of Proceedings of Machine Learning Research, 10–15 Jul 2018, pp. 3150–3158.

[90] L. T. Liu, M. Simchowitz, and M. Hardt, *The implicit fairness criterion of unconstrained learning*, in International Conference on Machine Learning, PMLR, 2019, pp. 4051–4060.

[91] P. K. Lohia, K. N. Ramamurthy, M. Bhide, D. Saha, K. R. Varshney, and R. Puri, *Bias mitigation post-processing for individual and group fairness*, in Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp), IEEE, 2019, pp. 2847–2851.

[92] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, *The variational fair autoencoder*, arXiv preprint arXiv:1511.00830, (2015).

[93] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, *Learning word vectors for sentiment analysis*, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, June 2011, Association for Computational Linguistics, pp. 142–150.

[94] D. Madras, E. Creager, T. Pitassi, and R. Zemel, *Learning adversarially fair and transferable representations*, arXiv preprint arXiv:1802.06309, (2018).

[95] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, *Towards deep learning models resistant to adversarial attacks*, arXiv preprint arXiv:1706.06083, (2017).

[96] M. Mahsereci, L. Balles, C. Lassner, and P. Hennig, *Early stopping without a validation set*, arXiv preprint arXiv:1703.09580, (2017).

[97] A. Majeed and S. Lee, *Anonymization techniques for privacy preserving data publishing: A comprehensive survey*, IEEE Access, (2020).

[98] D. Mandal, S. Deng, S. Jana, J. Wing, and D. J. Hsu, *Ensuring fairness beyond the training data*, in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds., vol. 33, Curran Associates, Inc., 2020, pp. 18445–18456.

[99] D. McNamara, C. S. Ong, and R. C. Williamson, *Costs and benefits of fair representation learning*, in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 263–270.

[100] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, *A survey on bias and fairness in machine learning*, ACM Computing Surveys (CSUR), 54 (2021), pp. 1–35.

[101] N. Mehrabi, M. Naveed, F. Morstatter, and A. Galstyan, *Exacerbating algorithmic bias through fairness attacks*, arXiv preprint arXiv:2012.08723, (2020).

[102] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, *Deepfool: a simple and accurate method to fool deep neural networks*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2574–2582.

[103] V. Nanda, S. Dooley, S. Singla, S. Feizi, and J. P. Dickerson, *Fairness through robustness: Investigating robustness disparity in deep learning*, in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 466–477.

[104] V. Noroozi, S. Bahaadini, S. Sheikhi, N. Mojab, and P. S. Yu, *Leveraging semi-supervised learning for fairness using neural networks*, arXiv preprint arXiv:1912.13230, (2019).

[105] Z. OBERMEYER, B. POWERS, C. VOGELI, AND S. MULLAINATHAN, *Dissecting racial bias in an algorithm used to manage the health of populations*, Science, 366 (2019), pp. 447–453.

[106] T. PANCH, H. MATTIE, AND L. A. CELI, *The ‚Äúinconvenient truth‚Äù about ai in healthcare*, NPJ digital medicine, 2 (2019), pp. 1–3.

[107] M. PANNEKOEK AND G. SPIGLER, *Investigating trade-offs in utility, fairness and differential privacy in neural networks*, arXiv preprint arXiv:2102.05975, (2021).

[108] N. PAPERNOT, M. ABADI, U. ERLINGSSON, I. GOODFELLOW, AND K. TALWAR, *Semi-supervised knowledge transfer for deep learning from private training data*, arXiv preprint arXiv:1610.05755, (2016).

[109] A. PASZKE, S. GROSS, F. MASSA, A. LERER, J. BRADBURY, G. CHANAN, T. KILLEEN, Z. LIN, N. GIMELSHEIN, L. ANTIGA, ET AL., *Pytorch: An imperative style, high-performance deep learning library*, in Advances in neural information processing systems, 2019, pp. 8026–8037.

[110] J. PENNINGTON, R. SOCHER, AND C. D. MANNING, *Glove: Global vectors for word representation*, in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[111] G. PLEISS, M. RAGHAVAN, F. WU, J. KLEINBERG, AND K. Q. WEINBERGER, *On fairness and calibration*, arXiv preprint arXiv:1709.02012, (2017).

[112] L. PRECHELT, *Automatic early stopping using cross validation: quantifying the criteria*, Neural Networks, 11 (1998), pp. 761–767.

[113] D. PUJOL, R. MCKENNA, S. KUPPAM, M. HAY, A. MACHANAVAJJHALA, AND G. MIKLAU, *Fair decision making using privacy-protected data*, in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 189–199.

[114] M. RIGAKI AND S. GARCIA, *A survey of privacy attacks in machine learning*, arXiv preprint arXiv:2007.07646, (2020).

[115] B. Rodríguez-Gálvez, R. Thobaben, and M. Skoglund, *A variational approach to privacy and fairness*, arXiv preprint arXiv:2006.06332, (2020).

[116] C. Rösner and M. Schmidt, *Privacy Preserving Clustering with Constraints*, in 45th International Colloquium on Automata, Languages, and Programming (ICALP 2018), vol. 107, 2018, pp. 96:1–96:14.

[117] A. Ruoss, M. Balunovic, M. Fischer, and M. Vechev, *Learning certified individually fair representations*, in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds., vol. 33, Curran Associates, Inc., 2020, pp. 7584–7596.

[118] M. Schmidt, C. Schwiegelshohn, and C. Sohler, *Fair coresets and streaming algorithms for fair k-means clustering*, CoRR, abs/1812.10854 (2018).

[119] U. S. Shanthamallu, A. Spanias, C. Tepedelenlioglu, and M. Stanley, *A brief survey of machine learning methods and their sensor and iot applications*, in 2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA), IEEE, 2017, pp. 1–8.

[120] S. Sharma, A. H. Gee, D. Paydarfar, and J. Ghosh, *Fair-n: Fair and robust neural networks for structured data*, arXiv preprint arXiv:2010.06113, (2020).

[121] X. Shen, S. Diamond, Y. Gu, and S. Boyd, *Disciplined convex-concave programming*, in 2016 IEEE 55th Conference on Decision and Control (CDC), IEEE, 2016, pp. 1009–1014.

[122] S. H. Silva and P. Najafirad, *Opportunities and challenges in deep learning adversarial robustness: A survey*, arXiv preprint arXiv:2007.00753, (2020).

[123] A. Solanas, C. Patsakis, M. Conti, I. S. Vlachos, V. Ramos, F. Falcone, O. Postolache, P. A. Pérez-Martínez, R. Di Pietro, D. N. Perrea, et al., *Smart health: A context-aware health paradigm within smart cities*, IEEE Communications Magazine, 52 (2014), pp. 74–81.

[124] J. Song, P. Kalluri, A. Grover, S. Zhao, and S. Ermon, *Learning controllable fair representations*, in Proceedings of the 22nd International Conference

on Artificial Intelligence and Statistics (AISTATS) 2019,, vol. 89, 16–18 Apr 2019, pp. 2164–2173.

[125] S. SONG, K. CHAUDHURI, AND A. D. SARWATE, *Stochastic gradient descent with differentially private updates*, in 2013 IEEE Global Conference on Signal and Information Processing, 2013, pp. 245–248.

[126] ——, *Stochastic gradient descent with differentially private updates*, in 2013 IEEE Global Conference on Signal and Information Processing, 2013, pp. 245–248.

[127] H. SURESH AND J. V. GUTTAG, *A framework for understanding unintended consequences of machine learning*, arXiv preprint arXiv:1901.10002, (2019).

[128] C. SZEGEDY, W. ZAREMBA, I. SUTSKEVER, J. BRUNA, D. ERHAN, I. GOODFELLOW, AND R. FERGUS, *Intriguing properties of neural networks*, in International Conference on Learning Representations, 2014.

[129] T. TANAY AND L. GRIFFIN, *A boundary tilting persepective on the phenomenon of adversarial examples*, arXiv preprint arXiv:1608.07690, (2016).

[130] B. USTUN, Y. LIU, AND D. PARKES, *Fairness without harm: Decoupled classifiers with preference guarantees*, in International Conference on Machine Learning, PMLR, 2019, pp. 6373–6382.

[131] N. VIGDOR, *Apple card investigated after gender discrimination complaints*, The New York Times, (2019).

[132] U. VON LUXBURG, *A tutorial on spectral clustering*, Statistics and computing, 17 (2007), pp. 395–416.

[133] F. WANG AND C. ZHANG, *Label propagation through linear neighborhoods*, IEEE Transactions on Knowledge and Data Engineering, 20 (2007), pp. 55–67.

[134] Y. WEN, S. LI, AND K. JIA, *Towards understanding the regularization of adversarial robustness on neural networks*, in International Conference on Machine Learning, PMLR, 2020, pp. 10225–10235.

[135] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro, *Learning non-discriminatory predictors*, in Conference on Learning Theory, 2017, pp. 1920–1953.

[136] Y. Wu, L. Zhang, X. Wu, and H. Tong, *Pc-fairness: A unified framework for measuring causality-based fairness*, in Advances in Neural Information Processing Systems, 2019, pp. 3404–3414.

[137] ——, *Pc-fairness: A unified framework for measuring causality-based fairness*, in Advances in Neural Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds., vol. 32, Curran Associates, Inc., 2019.

[138] D. Xu, W. Du, and X. Wu, *Removing disparate impact of differentially private stochastic gradient descent on model accuracy*, arXiv preprint arXiv:2003.03699, (2020).

[139] D. Xu, S. Yuan, and X. Wu, *Achieving differential privacy and fairness in logistic regression*, in Companion Proceedings of The 2019 World Wide Web Conference, 2019, pp. 594–599.

[140] J. Yang and Y. Li, *Differentially private feature selection*, in 2014 International Joint Conference on Neural Networks (IJCNN), IEEE, 2014, pp. 4182–4189.

[141] M. Yang, T. Zhu, Y. Xiang, and W. Zhou, *Density-based location preservation for mobile crowdsensing with differential privacy*, Ieee Access, 6 (2018), pp. 14779–14789.

[142] Q. Ye, H. Hu, X. Meng, and H. Zheng, *Privkv: Key-value data collection with local differential privacy*, in 2019 IEEE Symposium on Security and Privacy (SP), IEEE, 2019, pp. 317–331.

[143] C. Yin, J. Xi, R. Sun, and J. Wang, *Location privacy protection based on differential privacy strategy for big data in industrial internet of things*, IEEE Transactions on Industrial Informatics, 14 (2017), pp. 3628–3636.

[144] L. Yu, L. Liu, C. Pu, M. E. Gursoy, and S. Truex, *Differentially private model publishing for deep learning*, in 2019 IEEE Symposium on Security and Privacy (SP), IEEE, 2019, pp. 332–349.

[145] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, *Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment*, in Proceedings of the 26th International Conference on World Wide Web, 2017, pp. 1171–1180.

[146] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, *Fairness Constraints: Mechanisms for Fair Classification*, in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, vol. 54, 20–22 Apr 2017, pp. 962–970.

[147] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, *Learning fair representations*, in International Conference on Machine Learning, 2013, pp. 325–333.

[148] B. H. Zhang, B. Lemoine, and M. Mitchell, *Mitigating unwanted biases with adversarial learning*, in Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 335–340.

[149] T. Zhang, T. Zhu, J. Li, M. Han, W. Zhou, and P. Yu, *Fairness in semi-supervised learning: Unlabeled data help to reduce discrimination*, IEEE Transactions on Knowledge and Data Engineering, (2020).

[150] Z. Zhang, Y. Song, and H. Qi, *Age progression/regression by conditional adversarial autoencoder*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5810–5818.

[151] H. Zhao and G. Gordon, *Inherent tradeoffs in learning fair representations*, in Advances in neural information processing systems, 2019, pp. 15675–15685.

[152] J. Zhao, T. Jung, Y. Wang, and X. Li, *Achieving differential privacy of data disclosure in the smart grid*, in IEEE INFOCOM 2014-IEEE Conference on Computer Communications, IEEE, 2014, pp. 504–512.

[153] T. Zhu, G. Li, W. Zhou, and S. Y. Philip, *Differentially private data publishing and analysis: A survey*, IEEE Transactions on Knowledge and Data Engineering, 29 (2017), pp. 1619–1638.

[154] T. Zhu and S. Y. Philip, *Applying differential privacy mechanism in artificial intelligence*, in 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), IEEE, 2019, pp. 1601–1609.

[155] T. Zhu, P. Xiong, G. Li, and W. Zhou, *Correlated differential privacy: Hiding information in non-iid data set*, IEEE Transactions on Information Forensics and Security, 10 (2015), pp. 229–242.