Advances in Machine Learning Safety

by

Songkai Xue

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2024

Doctoral Committee:

Associate Professor Yuekai Sun, Chair
Professor Qiaozhu Mei
Professor Kerby Shedden
Professor Ji Zhu

Songkai Xue
sxue@umich.edu
ORCID iD: 0009-0003-6534-1180

© Songkai Xue 2024
All Rights Reserved

To my parents.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Yuekai Sun. Taking me on as your student was one of the most fortunate events that could have ever occurred to me. You introduced me to the fascinating research field of machine learning fairness, as well as the broader research field of machine learning safety. We have been working together for over five years. During these years, you have been an advisor, a collaborator, a good friend, and a role model. You always encourage me, inspire me, and finally raise me to the level that is best for my personal development. Thank you, Yuekai!

Secondly, I would like to thank Kerby Shedden and Ji Zhu. During my time at the department, I have been interacting with you frequently. We collaborated on teaching, research, or both. Over the years, you have witnessed my progress and provided me with generous advice in both work and life. Thank you, Kerby and Ji.

In addition, I would also like to thank Qiaozhu Mei for taking the time to serve on my committee. Qiaozhu's participation in my committee greatly broadens the audience for my work. Thank you, Qiaozhu.

Next, I would like to thank my friends and collaborators. Thank you for joining me on this journey: Bo, Cheng, Daniele, Fan, Felipe, Gang, Hu, Jiacheng, Jing, Jinming, Menghan, Mikhail, Pramit, Ruofei, Saptarshi, Seamus, Shihao, Subha, Weibin, Weijing, Xuefei, Yang, Yanxin, Yidan, Yuanzhi, and Yue.

Finally, I would like to thank my parents. Thank you for your unconditional love and support. This dissertation is dedicated to you. I love you, mom and dad!

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF FIGURES vii

LIST OF APPENDICES xi

ABSTRACT xii

CHAPTER

I. Introduction 1

II. Calibrated Data-Dependent Constraints with Exact Satisfaction Guarantees 5

2.1 Introduction 5
2.1.1 Related Work 8
2.2 Single Expected Value Constraint 10
2.2.1 Stochastic Approximation for (2.1.3) 14
2.3 Multiple Expected Value Constraints 15
2.4 Simulations 18
2.5 Application to Fair Machine Learning 20
2.5.1 A Two-stage Method for Unknown Active Set . . . 21
2.5.2 Proxy Dual Function for Non-differentiable Constraints 22
2.5.3 Adult Experiments 23
2.6 Summary and Discussion 24

III. The Cost of Noise in Fair Machine Learning 25

3.1 Introduction 25
3.1.1 Related Work 27

3.2 Problem Setup 28
3.2.1 Fair Machine Learning as Constrained Stochastic Optimization 28
3.2.2 Fair Machine Learning with Noisy Sensitive Attributes 30
3.2.3 Asymptotic Relative Efficiency 31
3.3 Noise Cost in Fair Machine Learning 33
3.3.1 Application 1: Fair ML with Private Sensitive Attributes 36
3.3.2 Application 2: Fair ML with Missing Sensitive Attributes 40
3.3.3 General Theory 42
3.4 Simulations 45
3.5 Real Data Experiments 46
3.6 Summary and Discussion 48

IV. Distributionally Robust Performative Prediction 50

4.1 Introduction 50
4.1.1 Related Work 51
4.2 Methodology 53
4.2.1 Performative Prediction Essentials 53
4.2.2 Distributionally Robust Performative Prediction . . 56
4.2.3 Generalization Principle of DRPO 57
4.2.4 Benefits of DRPO: A Toy Example 58
4.3 Theory 60
4.3.1 Strong Duality of DRPR 60
4.3.2 Worst-case Distribution Map 60
4.3.3 Excess Risk Bound 61
4.3.4 Generalized Excess Risk Bound 62
4.4 Algorithms 63
4.4.1 Performative Risk Minimization 63
4.4.2 Distributionally Robust Performative Risk Minimization 64
4.4.3 Tilted Performative Risk Minimization 65
4.4.4 Calibration of Critical Radius 66
4.5 Experiments 67
4.5.1 Strategic Classification with Misspecified Cost Function 68
4.5.2 Partially Identifiable Distribution Map 71
4.5.3 Fairness without Demographics 72
4.6 Summary and Discussion 74

V. Concluding Remarks and Future Directions 75

5.1 Calibrated Data-Dependent Constraints 75

5.2 Statistical Cost of Data Noise in Constrained Learning 76
5.3 Model Misspecification in Performative Prediction 77

APPENDICES 78

A. Appendix of Chapter II 79

A.1 Proofs of Theorem 2.2.1 and Corollary 2.2.1 79
A.2 Proof of Theorem 2.3.1 80
A.3 Proof of Corollary 2.3.1 89
A.4 Dual Ascent Algorithm for (2.3.2) 90
A.5 Simulations: Details and More 90
A.6 Application to ε-Equality of Opportunity 95
A.7 Two-stage Method for Unknown Active Set 95
A.8 Proxy Dual Ascent Algorithm for Non-differentiable Constraints 98
A.9 Real Data Experiments: Details and More 99
A.10 Dual Form of the Robust Constraint Function (2.2.5) 101
A.11 Experiments on Additional Baseline and Dataset 102

B. Appendix of Chapter III 107

B.1 Linear Fairness Constraint 107
B.2 Linear-fractional Fairness Constraint 112
B.3 Proof of Theorem 3.3.2 114
B.4 Multiple Demographic Groups 125
B.5 Missing Sensitive Attributes 127

C. Appendix of Chapter IV 128

C.1 Toy Example in Section 4.2.4 (Continued) 128
C.2 Deferred Proofs 131
C.3 Experimental Details and More Results 133
C.4 Extension to General φ-Divergence 138
C.5 Discussion of Algorithmic Convergence 142

BIBLIOGRAPHY 143

LIST OF FIGURES

Figure

II.1 Frequency versus limiting probability of constraint satisfaction of the first constraint (top left), the second constraint (top right), and both of the constraints (bottom). 19

II.2 Frequency of constraint satisfaction (left) and classification error rate (right) for different demographic parity tolerance $\varepsilon \in \{0.01,\ 0.02,\ 0.03,\ 0.04,\ 0.05\}$. Baseline (sample average approximation, SAA) and our methods (with nominal probability $0.60, 0.75, 0.90, 0.95$) are compared. 23

III.1 Asymptotic relative efficiency curve of γ for varying r_1 and r_2. . . . 39

III.2 Asymptotic relative efficiency curve of p for $\alpha = 0.5$ (top left), $\alpha = 1$ (top right), and $\alpha = 2$ (bottom). 42

III.3 Relative efficiency curves for $\pi_0 = 0.3$ (left) and $\pi_0 = 0.5$ (right). . . 46

III.4 Results of experiments with private sensitive attributes. Top left: boxplot of training signed constraint violation. Top right: boxplot of test signed constraint violation. Bottom left: density plot of test signed constraint violation. Bottom right: empirical and theoretical curves of relative efficiency. 47

III.5 Results of experiments with missing sensitive attributes. Empirical and theoretical curves of relative efficiency. 48

IV.1 Results of Experiment 4.5.1. Top left: performative risk incurred by the PO and the DRPO's with various radius ρ's. Top right: relative improvement in worst-case performative risk of the DRPO to the PO as the radius ρ increases, for different range of misspecification η's. Bottom: radius ρ versus estimated KL divergence between $\mathcal{D}_{\text{true}}(\theta_{\text{DRPO}}(\rho))$ and $\mathcal{D}(\theta_{\text{DRPO}}(\rho))$, where vertical bands indicate the calibrated radius ρ_{cal}'s. 69

IV.2 Histogram of performative loss under Experiment 4.5.1 with $\epsilon_{\text{true}} = 0.5$. Top left: histogram for the PO. Top right: histogram for the DRPO with $\rho = 0.02$. Bottom: hitogram for the DRPO with $\rho = 0.04$. 70

IV.3 Results of Experiment 4.5.2. Top left: performative risk incurred by the PO and the TPO's with various tilt α's. Top right: relative improvement in worst-case performative risk of the TPO to the PO as the tilt α increases, for different range of misspecification η's. Bottom: the correspondence relationship between the radius ρ and the (inverse of) optimal dual variable $\mu^\star$. 72

IV.4 Results of Experiment 4.5.3. Performative risk of the population, the majority, and the minority, as the tilt α increases. The vertical band indicates the calibrated tilt α_{cal}'s. 73

A.1 Frequency versus limiting probability of constraint satisfaction of the first constraint (left), the second constraint (middle), and both of the constraints (right). 93

A.2 Frequency versus limiting probability of constraint satisfaction. . . . 94

A.3 Frequency of constraint satisfaction (left) and classification error rate (right) for different false positive rate parity tolerance $\varepsilon \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$. Baseline (sample average approximation, SAA) and our methods (with nominal probability $0.60, 0.75, 0.90, 0.95$) are compared. 101

A.4 Frequency of constraint satisfaction (left) and classification error rate (right) for different demographic parity tolerance $\varepsilon \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$. Baseline (sample average approximation, SAA), our methods (with nominal probability $0.60, 0.75, 0.90, 0.95$), and two-dataset approach (Cotter et al., 2019a) are compared. 103

A.5 Frequency of constraint satisfaction (left) and classification error rate (right) for different false positive rate parity tolerance $\varepsilon \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$. Baseline (sample average approximation, SAA), our methods (with nominal probability $0.60, 0.75, 0.90, 0.95$), and two-dataset approach (Cotter et al., 2019a) are compared. 104

A.6 Frequency of constraint satisfaction (left) and classification error rate (right) for different demographic parity tolerance $\varepsilon \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$. Our methods (with nominal probability 0.75, 0.90, 0.95), sample average approximation (SAA) and two-dataset approach (Cotter et al., 2019a) are compared. 105

A.7 Frequency of constraint satisfaction (left) and classification error rate (right) for different true positive rate parity tolerance $\varepsilon \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$. Our methods (with nominal probability 0.75, 0.90,0.95), sample average approximation (SAA) and two-dataset approach (Cotter et al., 2019a) are compared. 106

B.1 Asymptotic relative efficiency curve of γ for varying ratios of π_0 to π_1 and $\mathrm{Var}[g(\theta^\star; X, Y)|A = 0]$ to $\mathrm{Var}[g(\theta^\star; X, Y)|A = 1]$. 111

B.2 Asymptotic relative efficiency curve of γ for varying ratio of $\mathbb{P}(Y = 1, A = 0)$ to $\mathbb{P}(Y = 1, A = 1)$. 114

C.1 Additional results of Experiment 4.5.1. Left: performative balanced error rate incurred by the PO and the DRPO's with various radius ρ's. Right: relative improvement in worst-case performative balanced error rate of the DRPO to the PO as the radius ρ increases, for different range of misspecification η's. 135

C.2 Additional results of Experiment 4.5.2. Left: performative balanced error rate incurred by the PO and the TPO's with various tilt α's. Right: relative improvement in worst-case performative balanced error rate of the TPO to the PO as the tilt α increases, for different range of misspecification η's. 136
C.3 Additional results of Experiment 4.5.3. Performative accuracy of the population, the majority, and the minority, as the tilt α increases. . 138

LIST OF APPENDICES

Appendix

A. Appendix of Chapter II 79

B. Appendix of Chapter III 107

C. Appendix of Chapter IV 128

ABSTRACT

As more and more machine learning (ML) and artificial intelligence (AI) systems are used in practical settings, it is essential that they exhibit responsible, trustworthy, and safe behavior. This dissertation contributes to machine learning safety by focusing on the key aspects of alignment, data quality, distributional robustness, and their intersections. It is largely motivated by concerns about discrepancies between societal expectations and the performance of end-to-end ML systems, data quality degradation caused by privacy protocols or data noise, and distribution shifts posed by dynamic interactions between ML models and their environments.

This dissertation zeroes in on three problems that arise in machine learning safety. They are:

1. calibrated data-dependent constraints: guarantee real-world goals are satisfied with a user-prescribed probability at test time;

2. statistical cost of data noise in constrained learning: quantify the effect of sensitive attribute noise on the generalizability of (fairness) constraints;

3. model misspecification in performative prediction: design a robust learning framework to better approximate the performative optimum.

By carefully employing a blend of tools from distributionally robust optimization, stochastic optimization, and statistical inference, this dissertation tackles these pressing issues in ML safety. The methodologies and insights provided herein lay the groundwork toward a more ethical and effective deployment of ML and AI technolo-

gies, contributing to ensuring that these systems align with human values and exhibit reliability against the complexities of real-world applications.

CHAPTER I

Introduction

In recent years, we have observed the swift progress and extensive implementation of machine learning (ML) and artificial intelligence (AI) technologies. Machine learning models are being used more and more in high-stakes decision-making tasks, such as recruitment and hiring (Gonzalez et al., 2019), recidivism prediction (Angwin et al., 2022), and healthcare (Davenport and Kalakota, 2019). Artificial intelligence systems have exhibited superintelligent capabilities in specific domains, such as excelling at the board game Go (AlphaGo, Silver et al. (2016)), automatically solving the protein-folding problem (AlphaFold, Jumper et al. (2021)), and effectively operating chatbots (GPT-4, Achiam et al. (2023)). Although these technologies enhance human productivity and augment people's abilities, they also bring about substantial risks. Malicious use of these technologies can lead to harmful outcomes, including biased decision-making (Dwork et al., 2012) and information manipulation (Goodfellow et al., 2014), exemplifying the occurrence referred to as ML/AI risks. This form of risk commonly arises when a system's behavior deviates unexpectedly from the intended design during the deployment of machine learning and artificial intelligence systems (Bengio et al., 2023). Due to the profound yet possibly dangerous impact of these systems, *machine learning safety* has become a critical field of study (Amodei et al., 2016). It is important to develop new theories and methods that can effectively

anticipate and manage these risks, ensuring the safe behavior of ML/AI systems in real-world applications.

Machine learning safety is an umbrella term that includes a broad range of topics related to evaluating and mitigating the potential harms arising from machine learning risks (Hendrycks et al., 2021; Mohseni et al., 2022). This dissertation does not attempt to cover all possible aspects but addresses three concrete problems concerning the intersections of *alignment*, *data quality*, and *distributional robustness*. Now we motivate these concrete problems from specific ML safety *risks* and outline our contributions.

Alignment. End-to-end machine learning systems may not perform in the way that society expects (Gabriel, 2020). Misaligned ML models can result in *societal risks*, such as inequality and unfairness (Angwin et al., 2022), by failing to accurately represent and uphold the ethical standards and social norms expected by communities. To ensure that ML systems properly align with human objectives and societal values, machine learning models are trained with data-dependent constraints (Goh et al., 2016). Such constraints often arise as empirical versions of expected value constraints that enforce fairness (Zafar et al., 2019), stability (Milani Fard et al., 2016), or other real-world goals. In Chapter II, we consider the task of training machine learning models with data-dependent constraints. We reformulate data-dependent constraints so that they are calibrated: enforcing the reformulated constraints guarantees that their expected value counterparts are satisfied with a user-prescribed probability. The resulting optimization problem is amendable to standard stochastic optimization algorithms, and we demonstrate the efficacy of our method on a fairness-sensitive classification task where we wish to guarantee the classifier's fairness (at test time).

Data quality. Data is central to the training of a machine learning model, and its quality (Gudivada et al., 2017) can be affected intentionally or unintentionally by data privacy protocols (Dwork, 2006) or noise in collected data (Natarajan et al., 2013).

Training a machine learning model with imperfect or deficient data *risks amplifying the uncertainty* of the resulting model (Hüllermeier and Waegeman, 2021). A common task in fair machine learning is training ML models that preserve certain summary statistics across subpopulations defined by sensitive attributes (Mitchell et al., 2021). However, access to such sensitive attributes in training data is usually restricted and the learner must therefore rely on noisy proxies for the sensitive attributes (Mozannar et al., 2020; Wang et al., 2020). In Chapter III, we study the effect of a privacy mechanism and a missing data mechanism that obfuscate the sensitive attributes from the learner on the fairness of the resulting classifier. We propose a statistically principled metric to characterize the cost of noise on fairness, and show that this cost manifests as a decline in the generalizability of fairness constraints.

Distributional robustness. Dynamic interactions between ML models and their environments bring about distribution shifts in the social world (Wagner et al., 2021). Failure to capture long-term distribution shifts may expose systems to *credibility risk*, where the consequences of errors become compounded over time, affecting the reliability and fairness of ML models in dynamic environments (D'Amour et al., 2022). Performative prediction (Perdomo et al., 2020) is an emergent learning paradigm that aims to model such scenarios where predictive outcomes subsequently influence the very systems they target. The pursuit of a performative optimum (PO)—minimizing performative risk—is generally reliant on modeling of the distribution map (Mendler-Dünner et al., 2020), which characterizes how a deployed ML model alters the data distribution. Unfortunately, inevitable misspecification of the distribution map can lead to a poor approximation of the true PO (Lin and Zrnic, 2023). In Chapter IV, we address this issue by introducing a novel framework of distributionally robust performative prediction and study a new solution concept termed as distributionally robust performative optimum (DRPO). We show provable guarantees for DRPO as a robust approximation to the true PO when the nominal distribution map is different

from the actual one. Moreover, distributionally robust performative prediction can be reformulated as an augmented performative prediction problem, enabling efficient optimization. The experimental results demonstrate that DRPO offers potential advantages over traditional PO approach when the distribution map is misspecified at either micro- or macro-level.

CHAPTER II

Calibrated Data-Dependent Constraints with Exact Satisfaction Guarantees

2.1 Introduction

In machine learning (ML) practice, accuracy is often only one of many training objectives. For example, algorithmic fairness considerations may require a credit scoring system to perform comparably on men and women. Here are a few other examples.

Churn rate and stability The churn rate of an ML model compared to another model is the fraction of samples on which the predictions of the two models differ (Goh et al., 2016; Milani Fard et al., 2016). In ML practice, one may wish to control the churn rate between a new model and its predecessor because a high churn rate can disorient users and downstream system components. One way of training models with small churn is to enforce a churn rate constraint during training.

Precision, recall, *etc.* Classification and information retrieval models must often balance precision and recall. To train such models, practitioners carefully trade off one metric for the other by optimizing for one metric subject to constraints on the other.

Resource constraints Practitioners sometimes wish to control how often a classifier predicts a certain class due to budget or resource constraints. For example, a company that uses ML to select customers for a targeted offer may wish to constrain the fraction of customers selected for the offer. Another prominent example of a stochastic optimization problem with resource constraints is the newsvendor problem, which we come back to in Section 2.4.

Unlike constraints on the structure of model parameters (*e.g.*, sparsity), the constraints encoding the preceding training objectives are *data-dependent*. This leads to the issue of *constraint generalization*: whether the constraints *generalize* out-of-sample. For example, if a classifier is trained to have comparable accuracy on two subpopulations in the training data, will it also have comparable accuracy on samples from the two subpopulations at test time?

In this chapter, we consider the out-of-sample generalization of *expected-value* constraints. To keep things simple, consider a stochastic optimization problem with a single *expected-value* constraint:

$$\theta^\star \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \mathbb{E}_{P_0}\big[f(\theta;Z)\big] = \int_{\mathcal{Z}} f(\theta;z)dP_0(z) \\ \text{subject to} & \mathbb{E}_{P_0}\big[g(\theta;Z)\big] = \int_{\mathcal{Z}} g(\theta;z)dP_0(z) \leq 0 \end{array} \right\}, \tag{2.1.1}$$

where Θ is a (finite-dimensional) parameter space, $f, g : \Theta \times \mathcal{Z} \to \mathbb{R}$ are (known) cost and constraint functions, and $Z \in \mathcal{Z}$ is a random variable that represents a sample. The distribution of Z is unknown, so we cannot solve (2.1.1) directly. Instead, we obtain IID training samples $\{Z_i\}_{i=1}^n$ from the true underlying distribution P_0 and solve the empirical version of (2.1.1):

$$\widehat{\theta}_n \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \frac{1}{n}\sum_{i=1}^n f(\theta;Z_i) \\ \text{subject to} & \frac{1}{n}\sum_{i=1}^n g(\theta;Z_i) \leq 0 \end{array} \right\}. \tag{2.1.2}$$

The estimator $\widehat{\theta}_n$ (of $\theta^\star$) is guaranteed to satisfy the empirical constraint, that is,

$$\frac{1}{n}\sum_{i=1}^{n} g(\widehat{\theta}_n; Z_i) \leq 0,$$

but it is unclear whether $\widehat{\theta}_n$ satisfies the actual (population) constraint

$$\mathbb{E}_{P_0}\big[g(\theta; Z)\big] \leq 0.$$

As we shall see, under standard assumptions on (2.1.1), $\widehat{\theta}_n$ only satisfies the actual constraint with probability approaching $\frac{1}{2}$ (see corollary 2.2.1). This is especially problematic for constraints that encode algorithmic fairness goals. For example, the 80% rule published by the US Equal Employment Opportunity Commission, interpreted in the machine learning context, requires the rate at which a classifier predicts the advantaged label in minority groups to be at least 80% of the rate at which the classifier predicts the advantaged label in the majority group (Barocas et al., 2023).

In this chapter, we propose a distributionally robust version of (2.1.2) that *guarantees* the actual constraint $\mathbb{E}_{P_0}\big[g(\theta; Z)\big] \leq 0$ will be satisfied with probability $1-\alpha$:

$$\widehat{\theta}_n \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \frac{1}{n}\sum_{i=1}^{n} f(\theta; Z_i) \\ \text{subject to} & \sup_{P: D_\varphi(P\|\widehat{P}_n) \leq \frac{\rho_\alpha}{n}} \mathbb{E}_P\big[g(\theta; Z)\big] \leq 0 \end{array} \right\}, \tag{2.1.3}$$

where D_φ is a φ-divergence (see Section 2.2 for details), $\widehat{P}_n$ is the empirical distribution of the training samples, and $\sqrt{\rho_\alpha}$ is the $1-\alpha$ quantile of a standard normal random variable. More concretely, we show that $\widehat{\theta}_n$ achieves *asymptotically exact constraint satisfaction*

$$\lim_{n\to\infty} \mathbb{P}\Big\{\mathbb{E}_{P_0}\big[g(\widehat{\theta}_n; Z)\big] \leq 0\Big\} = 1-\alpha. \tag{2.1.4}$$

Here the inner expectation is with respect to Z; the outer probability is with respect to the training samples $\{Z_i\}_{i=1}^n$. Three desirable properties of (2.1.3) are

1. **exact constraint satisfaction:** If the actual probability of constraint satisfaction exceeds $1-\alpha$, then the method is too conservative. This may (unnecessarily) increase the cost of the model. By picking ρ_α in (2.1.3) carefully, constraints are satisfied with asymptotically exact probability $1-\alpha$.

2. **computationally efficient:** As we shall see, the computational cost of solving (2.1.3) is comparable to the cost of solving distributionally robust sample average approximation (SAA) problems.

3. **pivotal:** There are no nuisance parameters to estimate (*e.g.*, asymptotic variances) in (2.1.3). The user merely needs to look up the correct quantile of the standard normal distribution for their desired level of constraint generalization.

The rest of this chapter is organized as follows. In Section 2.2, we develop method, theory, and algorithm for stochastic optimization problems with single constraint. In Section 2.3, we extend our method, theory, and algorithm to stochastic optimization problems with multiple constraints. In Section 2.4, we validate our theory by simulating a resource-constrained newsvendor problem. In Section 2.5, we demonstrate the efficacy of our method by using it to train an algorithmically fair income classifier. In addition, we show how to apply our method to a fairness constrained learning problem and discuss two practical considerations for fair ML application scenarios. Finally, we summarize our work in Section 2.6 and point out an interesting avenue of future work.

2.1.1 Related Work

The closest work to our work is Lam (2019). They seek to pick a (data-dependent) *uncertainty set* $\mathcal{U}$ such that

$$\lim_{n\to\infty} \mathbb{P}\left\{\sup_\theta \left\{\mathbb{E}_{P_0}\left[g(\theta;Z)\right] - \sup_{P\in\mathcal{U}} \mathbb{E}_P\left[g(\theta;Z)\right]\right\} \le 0\right\} = 1-\alpha. \tag{2.1.5}$$

This condition is stronger than necessary: we only require

$$\lim_{n\to\infty} \mathbb{P}\left\{\mathbb{E}_{P_0}\left[g(\widehat{\theta}_n; Z)\right] - \sup_{P\in\mathcal{U}} \mathbb{E}_P\left[g(\widehat{\theta}_n; Z)\right] \leq 0\right\} = 1-\alpha \tag{2.1.6}$$

where $\widehat{\theta}_n$ is a (data-dependent) estimator (not necessarily (2.1.2) or (2.1.3)). Lam (2019) study (asymptotic) constraint satisfaction (2.1.4) for all deterministic objective functions (see Lam (2019), §1.1 for details). They advocate picking a KL divergence ball with radius that depends on the excursion probability of a certain χ^2 process.

Another closely related line of work is on data-splitting approaches for ensuring constraint generalization (Woodworth et al., 2017; Cotter et al., 2019a). At a high level, they split the training data into a training and validation subsets and use the validation subset to tune models trained on the training subset so that they satisfy the constraints. Although (computationally) simple and intuitive, their approach does not allow users to precisely control the constraint violation probability.

Lam (2019) is the latest in a line of work on distributionally robust optimization (DRO) that show the optimal value of DRO problems

$$\min_{\theta\in\Theta} \sup_{P\in\mathcal{U}} \mathbb{E}_P\left[g(\theta; Z)\right], \tag{2.1.7}$$

where $\mathcal{U}$ is a (data-dependent) uncertainty set of probability distributions, are upper confidence bounds for the optimal values of stochastic optimization problems. Common choices of uncertainty sets in DRO include uncertainty sets defined by moment or support constraints (Chen et al., 2007; Delage and Ye, 2010; Goh and Sim, 2010), φ-divergences (Ben-Tal et al., 2012; Lam and Zhou, 2015; Namkoong and Duchi, 2016), and Wasserstein distances (Shafieezadeh Abadeh et al., 2015; Blanchet et al., 2019; Mohajerin Esfahani and Kuhn, 2018; Lee and Raginsky, 2018; Sinha et al., 2017). This line of work is motivated by Owen's seminal work on empirical likelihood (Owen, 1990). In recent work, Lam and Zhou (2015) and Duchi et al. (2021) show that

the optimal value of DRO problems with empirical likelihood uncertainty sets leads to asymptotically exact upper confidence bounds for the optimal value of stochastic optimization problems (Duchi et al. (2021) consider more general φ-divergence uncertainty sets). Blanchet et al. (2019) establish similar coverage results for Wasserstein uncertainty sets.

Our work is also closely related to the work on the variance regularization properties of DRO (Namkoong and Duchi, 2016), which uses DRO to approximate the variance regularization cost function (see (2.2.4)). Gao et al. (2022) establish similar results for Wasserstein DRO. Lastly, we relate our work to the literature on chance constrained optimization (see Jiang and Guan (2016) and the references therein). The general goal of chance constrained optimization is to minimize a loss function subject to the probability of satisfying uncertain constraints is above a prescribed level. While our methods reformulate expected value constraints and we show that the solution of the reformulated problem enjoys an asymptotically exact probabilistic guarantees of constraint satisfaction. In addition, the data-dependent constraints in our work are also unknown in practice, which differs from the common setup in the chance constrained optimization literature.

2.2 Single Expected Value Constraint

We motivate (2.1.3) by considering a few alternatives. First, we note that the results later in this section show that (2.1.2) violates the actual constraint in (2.1.1) approximately half the time (see corollary 2.2.1). The most straightforward modification of (2.1.2) to ensure $\widehat{\theta}_n$ satisfies the (actual) constraint $\mathbb{E}_{P_0}\big[g(\theta; Z)\big] \leq 0$ is to add a "margin" in (2.1.3); *i.e.* enforce the constraint

$$\frac{1}{n}\sum_{i=1}^{n} g(\theta; Z_i) + \epsilon_n \leq 0 \tag{2.2.1}$$

in (2.1.2). If we pick the slack term ϵ_n such that

$$\mathbb{P}\left\{\sup_{\theta\in\Theta}\left\{\mathbb{E}_{P_0}\big[g(\theta;Z)\big]-\frac{1}{n}\sum_{i=1}^{n}g(\theta;Z_i)\right\}>\epsilon_n\right\}\leq\alpha,$$

then it is not hard to check that the resulting $\widehat{\theta}_n$ satisfies the (actual) constraint with probability greater than $1-\alpha$ (Wang and Ahmed, 2008; Luedtke and Ahmed, 2008). However, this approach is most likely conservative because the constraint is unnecessarily stringent for θ's such that $\frac{1}{n}\sum_{i=1}^{n}g(\theta;Z_i)$ is less variable. It is also not pivotal: ϵ_n is often set using bounds from (uniform) concentration inequalities, which typically depend on unknown problem parameters.

To relax the empirical constraint in a way that adapts to the variability of the empirical constraints, we replace the uniform margin in (2.2.1) with a parameter-dependent margin:

$$\frac{1}{n}\sum_{i=1}^{n}g(\theta;Z_i)+z_\alpha\frac{\widehat{\boldsymbol{\sigma}}(\theta)}{\sqrt{n}}\leq 0, \tag{2.2.2}$$

where z_α is the $1-\alpha$ quantile of a standard normal random variable and $\widehat{\boldsymbol{\sigma}}^2(\theta)$ is an estimate of the asymptotic variance of $g(\theta;Z)$. We recognize the (parameter-dependent) margin as (a multiple of) the standard error of the empirical constraint. It is possible to show that enforcing (2.2.2) achieves asymptotically exact constraint generalization (2.1.4) (Lam, 2019).

The main issue with this method is it is not amenable to standard stochastic optimization algorithms. In particular, even if the original constraint in (2.1.2) is convex, (2.2.2) is generally non-convex. Another issue is that it is not pivotal: the user must estimate the asymptotic variance of $g(\theta;Z)$.

To overcome these two issues, we consider a distributionally robust version of (2.1.2); *i.e.* enforcing

$$\sup_{P:D_\varphi(P\|\widehat{P}_n)\leq\frac{\rho_\alpha}{n}}\mathbb{E}_P\big[g(\theta;Z)\big]\leq 0, \tag{2.2.3}$$

where $D_\varphi(P\|Q) \triangleq \int \varphi(\frac{dP}{dQ})dQ$ is a φ-divergence. Common choices of φ include $\varphi(t) = (t-1)^2$ (which leads to the χ^2-divergence) and $\varphi(t) = -\log t + t - 1$ (which leads to the Kullback-Leibler divergence). Although there are many other choices for the uncertainty set in (2.2.3), we pick an φ-divergence ball because (i) (2.2.3) with an φ-divergence ball is asymptotically equivalent to (2.2.2):

$$\sup_{P:D_\varphi(P\|\widehat{P}_n)\leq \frac{\rho_\alpha}{n}} \mathbb{E}_P\big[g(\theta;Z)\big] \approx \frac{1}{n}\sum_{i=1}^{n} g(\theta;Z_i) + z_\alpha \frac{\widehat{\boldsymbol{\sigma}}(\theta)}{\sqrt{n}}, \tag{2.2.4}$$

and (ii) it leads to pivotal uncertainty sets. For theoretical analysis, we always use $\varphi(t) = (t-1)^2$ and χ^2-divergence in the remainder of this chapter.

Before we state the asymptotically exact constraint satisfaction property of (2.1.3) rigorously, we describe our assumptions on the problem.

1. **smoothness and concentration:** f and g are twice continuously differentiable with respect to θ, and $f(\theta^\star;Z)$, $\nabla f(\theta^\star;Z)$, $g(\theta^\star;Z)$, $\nabla g(\theta^\star;Z)$ are sub-Gaussian random variables.

2. **uniqueness:** the stochastic optimization problem with a single expected value constraint (2.1.1) has a unique optimal primal-dual pair $(\theta^\star,\lambda^\star)$, and $\theta^\star$ belongs to the interior of the compact set Θ.

3. **strict complementarity:** $\lambda^\star > 0$.

4. **positive definiteness:** The Hessian of the Lagrangian evaluated at $(\theta^\star,\lambda^\star)$ is positive definite.

The preceding assumptions are not the most general, but they are easy to interpret. The smoothness conditions on f and g with respect to θ, the concentration conditions of $f(\theta^\star;Z)$ and $g(\theta^\star;Z)$, and the uniqueness condition facilitate the use of standard tools from asymptotic statistics to study the large sample properties of the constraint value. The strict complementarity condition rules out problems in which

the constraint is extraneous; *i.e.* problems in which the unconstrained minimum coincides with the constrained minimum.

We are ready to state the asymptotically exact constraint satisfaction property of (2.1.3) rigorously. The main technical result characterizes the limiting distribution of the constraint value.

Theorem 2.2.1. *Let $\widehat{\theta}_n$ be an optimal solution of* (2.1.3) *converging in probability as $n \to \infty$ to $\theta^\star$. Under the standing assumptions, we have*

$$\sqrt{n}\left(\mathbb{E}_{P_0}\big[g(\widehat{\theta}_n; Z)\big] - \cancel{\mathbb{E}_{P_0}\big[g(\theta^\star; Z)\big]}\right) \xrightarrow{d} \mathcal{N}\left(-\sqrt{\rho_\alpha \operatorname{Var}_{P_0}[g(\theta^\star; Z)]}, \operatorname{Var}_{P_0}[g(\theta^\star; Z)]\right).$$

We translate this result on the constraint value to a result on constraint generalization.

Corollary 2.2.1. *Let $\sqrt{\rho_\alpha}$ be the $1-\alpha$ quantile of a standard normal random variable. Under the conditions of theorem 2.2.1, we have*

$$\lim_{n\to\infty} \mathbb{P}\left\{\mathbb{E}_{P_0}\big[g(\widehat{\theta}_n; Z)\big] \le 0\right\} = \mathbb{P}\left\{U \le \sqrt{\rho_\alpha}\right\} = 1 - \alpha,$$

where $U \sim \mathcal{N}(0,1)$ is a standard Gaussian random variable.

From Theorem 2.2.1 and Corollary 2.2.1 (see proofs in Appendix A.1), we find that

1. picking $\rho_\alpha = 0$ (*i.e.*, equivalently solving (2.1.2)) leads to a constraint violation probability that approaches $\frac{1}{2}$ in the large sample limit.

2. the relation between the mean and variance of the limiting distribution of the constraint value in Theorem 2.2.1 allows us to pick ρ_α in a pivotal way (*i.e.* does not depend on nuisance parameters).

2.2.1 Stochastic Approximation for (2.1.3)

In the rest of this section, we derive a stochastic optimization algorithm to solve (2.1.3) efficiently. As we shall see, the computational cost of this algorithm is comparable to the cost of solving a DRO problem. The key insight is that the robust constraint function has a dual form (see Appendix A.10):

$$\sup_{P:D_\varphi(P\|\widehat{P}_n)\le\rho}\mathbb{E}_P\big[g(\theta;Z)\big]=\inf_{\mu\ge0,\nu\in\mathbb{R}}\left\{\frac{1}{n}\sum_{i=1}^n\mu\varphi^*\left(\frac{g(\theta;Z_i)-\nu}{\mu}\right)+\mu\rho+\nu\right\}, \tag{2.2.5}$$

where $\varphi^*(s)\triangleq\sup_t\{st-\varphi(t)\}$ is the convex conjugate of φ. As we use χ^2-squared divergence and $\varphi(t)=(t-1)^2$, the corresponding $\varphi^*(s)=\frac{s^2}{4}+s$. The Lagrangian of (2.1.3) is

$$\begin{aligned}L(\theta,\lambda)&\triangleq\frac{1}{n}\sum_{i=1}^n f(\theta;Z_i)+\lambda\sup_{P:D_\varphi(P\|\widehat{P}_n)\le\frac{\rho_\alpha}{n}}\mathbb{E}_P\big[g(\theta;Z)\big]\\&=\frac{1}{n}\sum_{i=1}^n f(\theta;Z_i)+\lambda\inf_{\mu\ge0,\nu\in\mathbb{R}}\left\{\frac{1}{n}\sum_{i=1}^n\mu\varphi^*\left(\frac{g(\theta;Z_i)-\nu}{\mu}\right)+\mu\frac{\rho_\alpha}{n}+\nu\right\}.\end{aligned}$$

We see that evaluating the dual function $\inf_\theta L(\theta,\lambda)$ (at a fixed λ) entails solving a stochastic optimization problem that is suitable for stochastic approximation. This suggests a dual ascent algorithm for solving (2.1.3):

1. evaluate the dual function at λ_t by solving a stochastic optimization problem;
2. update λ_t with a dual ascent step.

We summarize this algorithm in Algorithm 1. The main cost of Algorithm 1 is incurred in the third line: evaluating the dual function. Fortunately, this step is suitable for stochastic approximation, so we can leverage recent advances in the literature to reduce the (computational) cost of this step. The total cost of this algorithm is comparable to that of distributionally robust optimization.

Algorithm 1 Dual ascent algorithm for (2.1.3)

1: **Input:** starting dual iterate $\lambda_0 \geq 0$
2: **repeat**
3: Evaluate dual function:

$$(\theta_t, \mu_t, \nu_t) \leftarrow \arg\min_{\theta, \mu \geq 0, \nu} \frac{1}{n} \sum_{i=1}^n f(\theta; Z_i) + \lambda_t \left\{ \frac{1}{n} \sum_{i=1}^n \mu \varphi^* \left(\frac{g(\theta; Z_i) - \nu}{\mu} \right) + \mu \frac{\rho_\alpha}{n} + \nu \right\}$$

4: Dual ascent update: $\lambda_{t+1} \leftarrow \left[\lambda_t + \eta_t \left\{ \frac{1}{n} \sum_{i=1}^n \mu_t \varphi^* \left(\frac{g(\theta_t; Z_i) - \nu_t}{\mu_t} \right) + \mu_t \frac{\rho_\alpha}{n} + \nu_t \right\} \right]_+$
5: **until** converged

2.3 Multiple Expected Value Constraints

In this section, we extend the results from the preceding section to stochastic optimization problems with multiple data-dependent constraints. Consider a stochastic optimization problem with K expected value constraints

$$\theta^\star \in \left\{ \begin{array}{ll} \arg\min_{\theta \in \Theta} & \mathbb{E}_{P_0}\big[f(\theta; Z)\big] \\ \text{subject to} & \big\{\mathbb{E}_{P_0}\big[g_k(\theta; Z)\big] \leq 0\big\}_{k=1}^K \end{array} \right\}, \tag{2.3.1}$$

Following the development in Section 2.2, we enforce the expected value constraints with robust versions of the sample average constraints:

$$\widehat{\theta}_n \in \left\{ \begin{array}{ll} \arg\min_{\theta \in \Theta} & \frac{1}{n} \sum_{i=1}^n f(\theta; Z_i) \\ \text{subject to} & \left\{\sup_{P: D_\varphi(P \| \widehat{P}_n) \leq \frac{\rho_k}{n}} \mathbb{E}_P\big[g_k(\theta; Z)\big] \leq 0\right\}_{k=1}^K \end{array} \right\}, \tag{2.3.2}$$

where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_K)^\top$ are uncertainty set radii for the constraints. There are other approaches to enforcing multiple constraints that result in constraint generalization; we focus on (2.3.2) here because it allows the user to adjust the constraint generalization probability for different constraints.

First, we extend Theorem 2.2.1 and Corollary 2.2.1 to problems with multiple (expected value) constraints. We assume

1. **smoothness and concentration:** for $k \in [K]$, f, g_k are twice continuously

differentiable with respect to θ, and $f(\theta^\star; Z), \nabla f(\theta^\star; Z),\ g_k(\theta^\star; Z), \nabla g_k(\theta^\star; Z)$ are sub-Gaussian random variables.

2. **uniqueness:** the stochastic optimization problem with K expected value constraints (2.3.1) has a unique optimal primal-dual pair $(\theta^\star, \boldsymbol{\lambda}^\star)$, and $\theta^\star$ belongs to the interior of the compact set Θ.

3. **strict complementarity:** $\boldsymbol{\lambda}^\star \in \operatorname{int}(\mathbb{R}^K_+)$, *i.e.*, each component of $\boldsymbol{\lambda}^\star$ is strictly positive.

4. **positive definiteness:** The Hessian of the Lagrangian evaluated at $(\theta^\star, \boldsymbol{\lambda}^\star)$ is positive definite.

The strict complementarity constraint seems especially strong here because it requires all the constraints to be active. It is possible (with extra notational overhead) to state the result in terms of just the active constraints. We refer to Section 2.5.1 for more information about the unknown active set. Further, as long as the sample size is large enough, the active constraints in (2.3.2) coincide with the active constraints in (2.3.1). To keep things simple, we assume all the constraints are active.

Theorem 2.3.1. *Let $\widehat{\theta}_n$ be an optimal solution of* (2.3.2) *converging in probability as $n \to \infty$ to $\theta^\star$. Under the standing assumptions, we have*

$$\sqrt{n}\begin{bmatrix} \mathbb{E}_{P_0}\big[g_1(\widehat{\theta}_n; Z)\big] \\ \vdots \\ \mathbb{E}_{P_0}\big[g_K(\widehat{\theta}_n; Z)\big] \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(-\begin{bmatrix} \sqrt{\rho_1 \operatorname{Var}_{P_0}[g_1(\theta^\star; Z)]} \\ \vdots \\ \sqrt{\rho_K \operatorname{Var}_{P_0}[g_K(\theta^\star; Z)]} \end{bmatrix}, \operatorname{Var}_{P_0}\begin{bmatrix} g_1(\theta^\star; Z) \\ \vdots \\ g_K(\theta^\star; Z) \end{bmatrix} \right).$$

Corollary 2.3.1. *Under the conditions of theorem 2.3.1, we have*

$$\lim_{n\to\infty} \mathbb{P}\left\{ \begin{bmatrix} \mathbb{E}_{P_0}\big[g_1(\widehat{\theta}_n; Z)\big] \\ \vdots \\ \mathbb{E}_{P_0}\big[g_K(\widehat{\theta}_n; Z)\big] \end{bmatrix} \in -\mathbb{R}^K_+ \right\} = \mathbb{P}\{\boldsymbol{U} \le \sqrt{\boldsymbol{\rho}}\},$$

where $\sqrt{\boldsymbol{\rho}} = (\sqrt{\rho_1}, \ldots, \sqrt{\rho_K})^\top$*, and* $\boldsymbol{U}$ *is a Gaussian random vector with mean zero and covariance*

$$\operatorname{Corr}_{P_0}\begin{bmatrix} g_1(\theta^\star; Z) \\ \vdots \\ g_K(\theta^\star; Z) \end{bmatrix} \triangleq D^{-\frac{1}{2}} \operatorname{Cov}_{P_0}\begin{bmatrix} g_1(\theta^\star; Z) \\ \vdots \\ g_K(\theta^\star; Z) \end{bmatrix} D^{-\frac{1}{2}}, \tag{2.3.3}$$
$$D \triangleq \operatorname{diag}\left(\{\operatorname{Var}_{P_0}[g_k(\theta^\star, Z)]\}_{k=1}^K\right).$$

From Theorem 2.3.1 and Corollary 2.3.1 (see proofs in Appendix A.2 and A.3), we find that the probability of constraint satisfaction decreases *exponentially* as the number of constraints increases. We also see that our method is no longer pivotal for multiple expected value constraints: the uncertainty set radii depends on the (unknown) correlation structure among the constraint values. Fortunately, it is not hard to estimate this correlation structure. The most straightforward way is with the empirical correlation matrix. Let $\widehat{\Sigma}_n$ be the empirical covariance matrix of the constraint values. The empirical correlation matrix is then given by $\widehat{R}_n \triangleq \operatorname{diag}(\widehat{\Sigma}_n)^{-\frac{1}{2}} \widehat{\Sigma}_n \operatorname{diag}(\widehat{\Sigma}_n)^{-\frac{1}{2}}$.

Finally, it is straightforward to extend the algorithm for solving (2.1.3) to (2.3.2). The Lagrangian of (2.3.2) is

$$\begin{aligned} L(\theta, \boldsymbol{\lambda}) &\triangleq \frac{1}{n}\sum_{i=1}^n f(\theta; Z_i) + \sum_{k=1}^K \lambda_k \sup_{P: D_\varphi(P\|P_n) \le \frac{\rho_k}{n}} \mathbb{E}_P\big[g_k(\theta; Z)\big] \\ &= \frac{1}{n}\sum_{i=1}^n f(\theta; Z_i) + \sum_{k=1}^K \lambda_k \inf_{\mu_k \ge 0, \nu_k \in \mathbb{R}} \left\{ \frac{1}{n}\sum_{i=1}^n \mu_k \varphi^*\left(\frac{g_k(\theta; Z_i) - \nu_k}{\mu_k}\right) + \mu_k \frac{\rho_k}{n} + \nu_k \right\}, \end{aligned}$$

where we recalled the dual form of the robust constraint function (2.2.5) in the second step. We see that evaluating the dual function $\inf_\theta L(\theta, \boldsymbol{\lambda})$ (at a fixed $\boldsymbol{\lambda}$) entails solving a stochastic optimization problem that is suitable for stochastic approximation. This suggests a similar dual ascent algorithm for solving (2.1.3); we skip the details here (see Algorithm 4 in Appendix A.4).

2.4 Simulations

We simulate the frequency of constraint satisfaction for the following multi-item newsvendor problem:

$$\begin{aligned} \max_{\theta\in\Theta} \quad & \mathbb{E}_{P_0}\left[p^\top \min\{Z,\theta\} - c^\top\theta\right] \\ \text{subject to} \quad & \mathbb{E}_{P_0}[(\|Z^{(1)}\|_2^2 - \|\theta^{(1)}\|_2^2)_+] \le \varepsilon_1 \\ & \mathbb{E}_{P_0}[(\|Z^{(2)}\|_2^2 - \|\theta^{(2)}\|_2^2)_+] \le \varepsilon_2 \end{aligned} \tag{2.4.1}$$

where $c \in \mathbb{R}_+^d$ is the manufacturing cost, $p \in \mathbb{R}_+^d$ is the sell price, $\theta \in \Theta = [0,100]^d$ is the number of items in stock, $Z \in \mathbb{R}^d$ is a random variable with probability distribution P_0 representing the demand, and there are d items in total. The distribution P_0 is unknown but we observe IID samples $Z_1, \ldots, Z_n$ from P_0. All of the items have been partitioned into two groups so that the corresponding demand and stock can be written as $Z = (Z^{(1)}, Z^{(2)})$ and $\theta = (\theta^{(1)}, \theta^{(2)})$. The constraints in the problem exclude stock levels that underestimate the demand too much for each group of items, where $\varepsilon_1, \varepsilon_2 > 0$ indicate tolerance level of such underestimation. The target of the problem is to maximize the profit while satisfying the constraints. It is easy to rewrite the maximization problem (2.4.1) as a minimization problem with expected value constraints in the form of (2.3.1) so that we can apply our method (2.3.2). We pick P_0 as multivariate Gaussian with independent components so that the two constraints are generally uncorrelated with each other (see Appendix A.5 for details).

Throughout the simulations, we solve (2.3.2) with $\boldsymbol{\rho} = (z_\alpha, z_\alpha)^\top$ for $\alpha \in \{0.4, 0.25, 0.1, 0.05, 0.005\}$. As suggested by our asymptotic theory in Section 2.3, the nominal probability of constraint satisfaction is $1-\alpha$ for each constraint and $(1-\alpha)^2$ for both constraints due to the independence setup.

In Figure II.1, we plot frequencies of constraint satisfaction for each constraint and both constraints, all of which are averaged over 1000 replicates. As the sample size

n grows, the frequency versus probability curve converges to the theoretical dashed line of limiting probability of constraint satisfaction, validating our theory in the large sample regime. For more simulations (*e.g.*, single constraint, two dependent constraints) we refer to Appendix A.5.

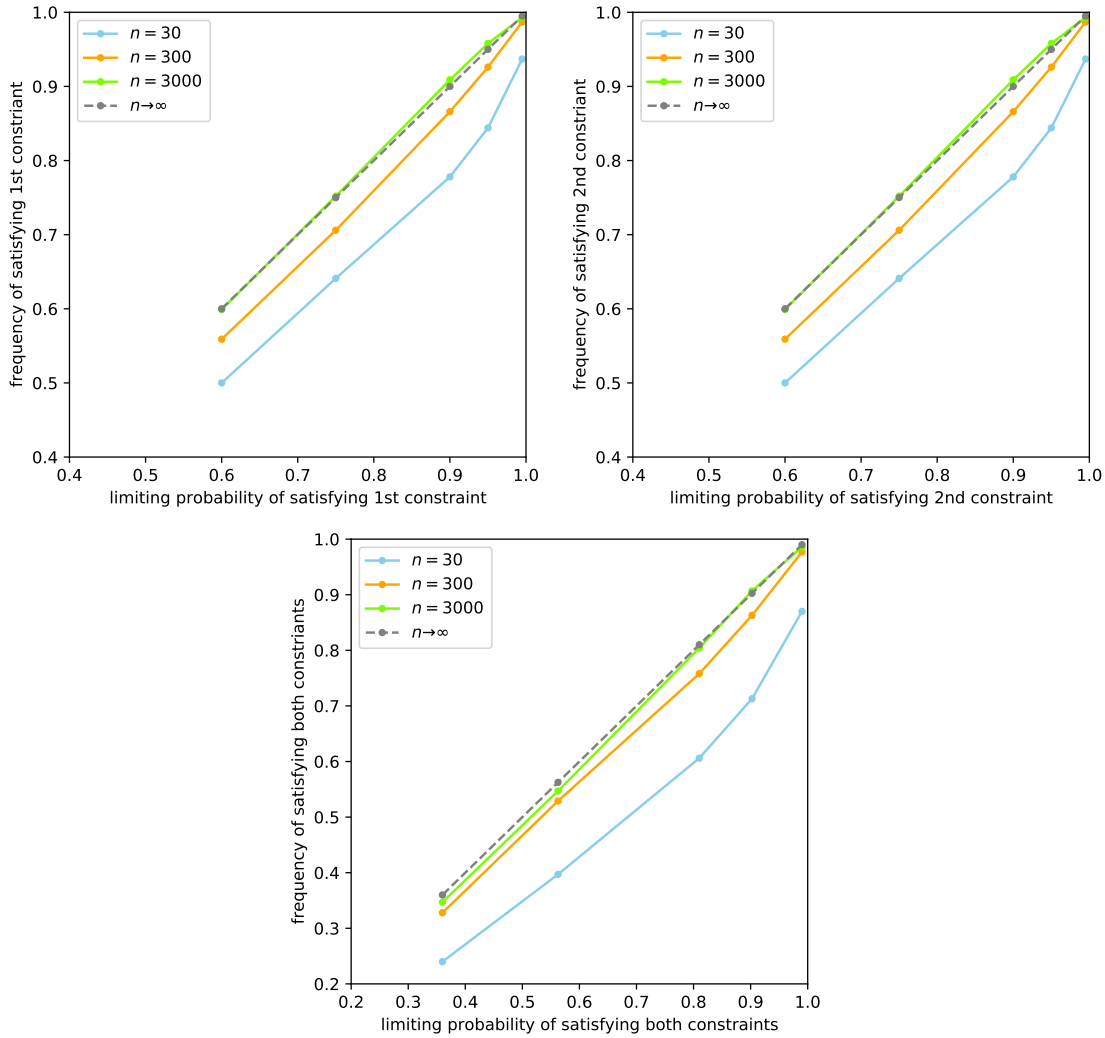

Figure II.1: Frequency versus limiting probability of constraint satisfaction of the first constraint (top left), the second constraint (top right), and both of the constraints (bottom).

2.5 Application to Fair Machine Learning

As ML models are deployed in high-stakes decision making and decision support roles, the fairness of the models has come under increased scrutiny. In response, there is a flurry of recent work on mathematical definitions of algorithmic fairness (Dwork et al., 2012; Hardt et al., 2016; Kusner et al., 2017) and algorithms to enforce the definitions (Agarwal et al., 2018; Cotter et al., 2019c; Yurochkin and Sun, 2020).

A prominent class of fairness definitions is *group fairness*; such definitions require equality of certain metrics (*e.g.* false/true positive rates) among demographic groups. For example, consider a fair binary classification problem. Let $\mathcal{X} \subset \mathbb{R}^d$ be the input space, $\{0, 1\}$ be the set of possible labels, and $\mathcal{A}$ be the set of possible values of the protected/sensitive attribute. In this setup, training and test examples are tuples of the form $(X, A, Y) \in \mathcal{X} \times \mathcal{A} \times \mathcal{Y}$, and a classifier is a map $f : \mathcal{X} \to \{0, 1\}$. A popular definition of algorithmic fairness for binary classification is *equality of opportunity* (Hardt et al., 2016).

Definition 2.5.1 (equality of opportunity)**.** Let $Y = 1$ be the advantaged label that is associated with a positive outcome and $\widehat{Y} \triangleq f(X)$ be the output of the classifier. Equality of opportunity entails $\mathbb{P}\{\widehat{Y} = 1 \mid A = a, Y = 1\} = \mathbb{P}\{\widehat{Y} = 1 \mid A = a', Y = 1\}$ for all $a, a' \in \mathcal{A}$.

Equality of opportunity, or true positive rate parity, means that the prediction $\widehat{Y} = h(X)$ conditioned on the advantaged label $Y = 1$ is statistically independent of the protected attribute A. Furthermore, an approximate version of equality of opportunity can be readily defined. We say that $\widehat{Y} = h(X)$ satisfies ε-*equality of opportunity* if $\mathbb{P}\{\widehat{Y} = 1 \mid A = a, Y = 1\} - \mathbb{P}\{\widehat{Y} = 1 \mid A = a', Y = 1\} \leq \varepsilon$ for for all $a, a' \in \mathcal{A}$. In this case, $\varepsilon > 0$ represents a practitioner's *tolerance* for fairness violations.

Given a parametric model space $\mathcal{H} = \{f_\theta(\cdot) : \theta \in \Theta\}$ and loss function ℓ :

$\Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, an in-processing fair ML routine is to minimize the (empirical) risk $\mathbb{E}\left[\ell(\theta; X, Y)\right]$ while satisfying some fairness constraints. Most commonly, definitions of group fairness (including equality of opportunity, demographic parity, and more) can be written as a special example of a general set of linear constraints (Agarwal et al., 2018, 2019) of the form $\mathbf{M}\boldsymbol{\mu}(\theta) \leq \boldsymbol{c}$, where matrix $\mathbf{M} \in \mathbb{R}^{K \times T}$ and vector $\boldsymbol{c} \in \mathbb{R}^K$ encode the constraints; $\boldsymbol{\mu}(\theta) : \Theta \to \mathbb{R}^T$ is a vector of (conditional) moments $\mu_t(\theta) = \mathbb{E}\left[h_t(X, A, Y, \theta) \mid \mathcal{E}_t\right]$ for $t \in [T]$; $g_t : \mathcal{X} \times \mathcal{A} \times \mathcal{Y} \times \Theta \to \mathbb{R}$; event $\mathcal{E}_t$ is defined with respect to (X, A, Y).

This framework fits to our methodology if we note that each (conditional) moment can be written as

$$\mu_t(\theta) = \frac{\mathbb{E}_{(X,A,Y)\sim P_0}\left[h_t(X, A, Y, \theta) \times \mathbf{1}\left\{\mathcal{E}_t(X, Y, A)\right\}\right]}{\mathbb{E}_{(X,A,Y)\sim P_0}\left[\mathbf{1}\left\{\mathcal{E}_t(X, Y, A)\right\}\right]}. \tag{2.5.1}$$

Here the indicator $\mathbf{1}\left\{\mathcal{E}_t\right\}$ takes value 1 if the event $\mathcal{E}_t$ happens, and 0 otherwise. Moreover, we use $\mathcal{E}_t(X, A, Y)$ to emphasize that $\mathcal{E}_t$ only depends on (X, Y, A) but not on θ in any way.

Note that (2.5.1) is a ratio of expected values, which is a non-linear statistical functional of P_0. To use our method, we first replace the denominator of $\mu_t(\theta)$ with an estimator, such as the unbiased estimator $\widehat{\mathbb{P}}(\mathcal{E}_t) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{\mathcal{E}_t(X_i, A_i, Y_i)\right\}$. The resulting plug-in estimation of $\mu_t(\theta)$ then becomes linear in P_0, allowing us to apply our method (see similar tricks in Cotter et al. (2019a)). We describe the application of our method to ε-equality of opportunity in Appendix A.6.

2.5.1 A Two-stage Method for Unknown Active Set

In practice, it is probable that only a subset of the constraints are active. Furthermore, we do not know beforehand whether or not a constraint is active in the true population problem. To handle this scenario, we propose a two-stage method:

1. At the first stage, we solve the sample average approximation (SAA) problem (2.3.2) with $\boldsymbol{\rho} = \mathbf{0}_K$. By doing so, we identify the active set of the SAA problem.

2. At the second stage, we solve (2.3.2) with $\boldsymbol{\rho}$ such that ρ_k is a positive number only if the k-th constraint, $k \in [K]$, was identified as active at the first stage.

In Appendix A.7, we show that the two-stage method also enjoys the calibration property (similar to Theorem 2.3.1 and Corollary 2.3.1) under standard assumptions (*i.e.*, strict complementarity). At a high level, the limiting probability of satisfying the true constraints depends solely on the correlation structure between active constraints and the uncertainty set radii for active constraints, as long as the SAA problem identifies active constraints with probability tending to 1.

2.5.2 Proxy Dual Function for Non-differentiable Constraints

Constraint functions in fair ML are often non-differentiable. For instance, fairness metrics are typically linear combinations of indicators that result in non-differentiable rate constraints (Cotter et al., 2019a,b,c). This prevents the use of any gradient-based optimization algorithms. Fortunately, only the dual function evaluation step in Algorithm 1 requires access to gradients. Therefore, we can modify the algorithm by: (1) introducing proxy dual function, which uses a differentiable surrogate $\tilde{g}$ instead of the non-differentiable g in the dual function evaluation step; (2) keeping g in the dual ascent step. For an indicator function $h(t) = \mathbf{1}\{t > 0\}$, one can replace it by sigmoidal function $h_1(t) = (1+e^{-at})^{-1}$ or hinge upper bound $h_2(t) = \max\{0, t+1\}$ to produce smooth surrogates for non-differentiable rate constraints (Davenport et al., 2010; Eban et al., 2017; Cotter et al., 2019b). We summarize the proxy dual ascent algorithm in Appendix A.8.

2.5.3 Adult Experiments

We compare the frequency of constraint satisfaction (at test time) of the sample average approximation and our methods with nominal probability $0.60, 0.75, 0.90, 0.95$ using the Adult dataset from UCI (Dua and Graff, 2017). The classification task is to predict whether an individual's income per year is higher than \$50K. The fairness goal is ε-demographic parity (ε-DP): $|\mathbb{P}(\widehat{Y} = 1 \mid A = 1) - \mathbb{P}(\widehat{Y} = 1 \mid A = 0)| \leq \varepsilon$, where $A = 1$ for male is the advantaged group and $A = 0$ for female is the disadvantaged group. We use a logistic regression model for classification and techniques in this Section for implementation.

In Figure II.2, we have line plots for frequency of constraint satisfaction and box plots for classification error rate, all of which are summarized over 100 replicates. The left panel shows that solving (2.3.1) directly leads to one half chance of constraint violation, while our method's constraint satisfaction frequency matches its nominal value. The price of a higher chance of test-time fairness satisfaction is an increase in classification error rate as shown in the right panel. From the baseline to 95% chance of fairness satisfaction, we basically trade off 2% increase in error rate. We refer to Appendix A.9 and A.11 for details and more experiments.

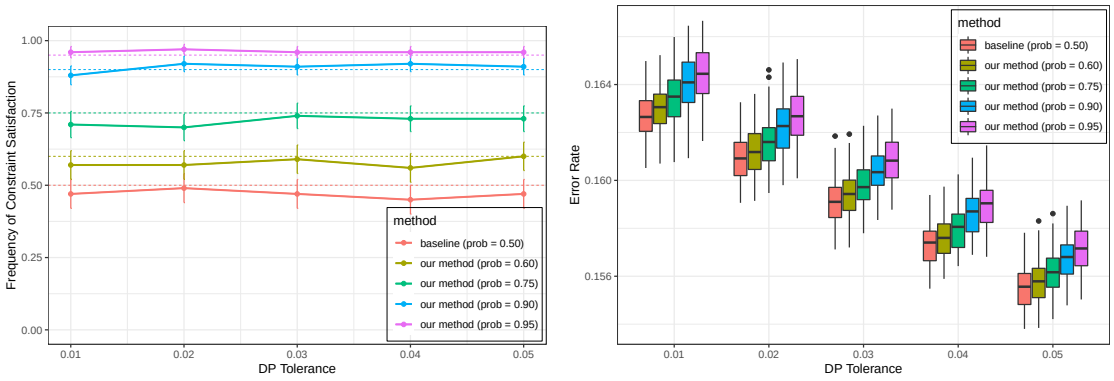

Figure II.2: Frequency of constraint satisfaction (left) and classification error rate (right) for different demographic parity tolerance $\varepsilon \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$. Baseline (sample average approximation, SAA) and our methods (with nominal probability $0.60, 0.75, 0.90, 0.95$) are compared.

2.6 Summary and Discussion

We explore the problem of exact constraint satisfaction probability in stochastic optimization with expected-value constraints. We propose a distributionally robust reformulation of data-dependent constraints and provide a theoretical guarantee of constraint satisfaction with an asymptotically exact probability specified by the user. For solving the reformulated problem, a scalable dual ascent algorithm and its variants are proposed. The computational cost of our algorithm is comparable to that of a standard distributionally robust optimization problem. Our theory on exact constraint satisfaction probability is validated via simulations on the resource-constrained newsvendor problem. The efficacy of our methods is empirically demonstrated on fair machine learning applications.

Some data-dependent constraints are by nature *non-linear* in the underlying probability measure. For example, (2.5.1) is a ratio of expected values. An intriguing direction for future research is to generalize the methods and theory developed in this work to constraints on non-linear functions of expected values. Such forms of constraints are known as *statistical functionals* in statistics literature (Fernholz, 2012). The non-linear dependence of the constraint function on the probability measure precludes the stochastic approximation as a general way of evaluating the dual function, as the constraint function no longer admits a dual form (2.2.5), calling for the development of a new algorithm.

CHAPTER III

The Cost of Noise in Fair Machine Learning

3.1 Introduction

The fairness of machine learning systems is gaining increasing attention in recent years. Among the numerous fairness objectives is ensuring that a machine learning model does not discriminate against subpopulations that are typically identified by sensitive attributes (*e.g.*, race, gender). When training a fair model and evaluating model bias, it is necessary to possess sensitive attributes; however, access to and use of such sensitive data is frequently prohibited by laws and regulations. Credit card companies, for instance, are not permitted to inquire about a person's race when they apply for credit, yet they must demonstrate that their decisions are not discriminatory (Chen et al., 2019).

Ideally, sensitive personal information should not be disclosed during the training of ML models. However, it is impossible to ensure exact notions of fairness (such as demographic parity or equality of opportunity) without any knowledge of the sensitive data. Fortunately, differential privacy (Dwork et al., 2006) is a promising workaround, which can offer a graceful compromise between privacy and utility. Mozannar et al. (2020) propose to release sensitive attributes in a locally differentially private way: adding noise to the sensitive data so that adversaries cannot infer any information with high confidence about a single record.

The advantage of the privacy mechanism proposed by Mozannar et al. (2020) is an invariance property: exact notions of fairness with regard to true sensitive attributes and noisy sensitive attributes are equivalent. An implication of the invariance property is that the optimal model of fairness can be learned at the population level. Nonetheless, it remains unclear what the precise statistical impact of privacy on fairness is.

Missingness is another source of noise in sensitive attributes. People may choose not to disclose their demographic identities during data collection due to privacy or other concerns, as is commonly the case with census and medical data. In medicine, it is reported that missing data in electronic health records is a key issue leading to the unfairness of ML models, which may increase healthcare inequities (Gianfrancesco et al., 2018).

In this chapter, we study the statistical cost of privacy and missingness on fairness in the task of learning fair ML models with noisy sensitive attributes. The main benefits of the developed theory are

1. **statistically principled:** We propose a statistically principled metric to characterize the cost of noise on fairness. A restricted notion of statistical efficiency *precisely quantifies* the privacy or missingness cost asymptotically.

2. **interpretable:** Privacy or missingness leads to a decline in the statistical efficiency. Such efficiency loss is interpretable: it explicitly depends on the privacy budget or the missing proportion, the subpopulation imbalance level, and few other problem-specific parameters.

The rest of this chapter is organized as follows. In Section 3.2, we formalize the problem setup, which consists of the constrained stochastic optimization problem for fair machine learning, the learning procedure of fair model using noisy sensitive attributes, and the definition of asymptotic relative efficiency in terms of fairness

violations. In Section 3.3, we develop theory for the noise cost under a single exact fairness constraint and then apply the main theory to a differential privacy mechanism and a missing data mechanism for releasing sensitive attributes. Also, we generalize this theory to the setting of multiple fairness constraints. By simulating a risk-parity regression problem in Section 3.4 and conducting a fairness-sensitive classification task on a real data in Section 3.5, we validate our theory and illustrate the utility of our tools. Finally, we summarize our work in Section 3.6 and point out an interesting avenue of future work.

3.1.1 Related Work

The interaction between fairness and privacy has been investigated from three perspectives: learning approximately fair models without sensitive attributes (Hashimoto et al., 2018; Lahoti et al., 2020), learning approximately fair models with wildly noisy sensitive attributes (Kallus et al., 2022; Awasthi et al., 2020; Wang et al., 2020), and learning exactly fair models with *structured* noisy sensitive attributes (Lamy et al., 2019; Mozannar et al., 2020). This chapter focuses on the third aspect.

The works that are most pertinent to ours are Lamy et al. (2019) and Mozannar et al. (2020). Lamy et al. (2019) assume that the sensitive attributes are subject to noise from the mutually contaminated learning model. Under such a structured noise mechanism, the noise rates can be consistently estimated, and when enforcing fairness with regard to noisy groups, scaling the fairness tolerance parameter more tightly is all that is required. Mozannar et al. (2020) suggest a differentially private model to release the sensitive attributes, which is a special type of the mutually contaminated learning model. Under such a designed noise mechanism, Mozannar et al. (2020) show that if the classifier is independent of the sensitive attributes, then exact fairness with regard to noisy sensitive attributes is equivalent to that with regard to true sensitive attributes. The idea of the equivalence can be found in Lamy et al. (2019) while

Mozannar et al. (2020) put it into a formal statement.

We basically study the statistical cost of noise induced by privacy or missingness of sensitive attirbutes on the generalizability of fairness when using Lamy et al. (2019)'s method under Mozannar et al. (2020)'s local differential privacy mechanism or missing at random mechanism (Allison, 2001).

3.2 Problem Setup

3.2.1 Fair Machine Learning as Constrained Stochastic Optimization

In-processing fair machine learning is typically a supervised learning problem with fairness constraints (Zafar et al., 2017; Agarwal et al., 2018). Such a problem can most often be formulated as a constrained stochastic optimization problem: (empirical) risk minimization subject to (empirical) fairness constraints.

Consider a fair binary classification problem. Let $\mathcal{X} \subset \mathbb{R}^d$ be the input space, $\mathcal{Y} = \{0, 1\}$ be the set of possible labels, and $\mathcal{A}$ be the set of possible values of the protected/sensitive attribute. In this setup, training and test examples are tuples of the form $(X, A, Y) \in \mathcal{X} \times \mathcal{A} \times \mathcal{Y}$, and a classifier is a map $f : \mathcal{X} \to \{0, 1\}$. Two popular definitions of algorithmic fairness for binary classification are *demographic parity* (Dwork et al., 2012) and *equality of opportunity* (Hardt et al., 2016).

Definition 3.2.1 (Demographic parity)**.** Let $\widehat{Y} = f(X)$ be the output of the classifier. Demographic parity entails $\mathbb{P}\{\widehat{Y} = 1 \mid A = a\} = \mathbb{P}\{\widehat{Y} = 1 \mid A = a'\}$ for all $a, a' \in \mathcal{A}$.

Demographic parity, also known as statistical parity, means that the prediction $\widehat{Y} = f(X)$ is statistically independent of the protected attribute A.

Definition 3.2.2 (Equality of opportunity)**.** Let $Y = 1$ be the advantaged label that is associated with a positive outcome and $\widehat{Y} = f(X)$ be the output of the classifier. Equality of opportunity entails $\mathbb{P}\{\widehat{Y} = 1 \mid A = a, Y = 1\} = \mathbb{P}\{\widehat{Y} = 1 \mid A = a', Y = 1\}$ for all $a, a' \in \mathcal{A}$.

Equality of opportunity, also known as true positive rate parity, means that the prediction $\widehat{Y} = f(X)$ conditioned on the advantaged label $Y = 1$ is statistically independent of the protected attribute A.

Given a parametric model space $\mathcal{H} = \{f_\theta(\cdot) : \theta \in \Theta\}$ and loss function $\ell : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ (where $\Theta \subset \mathbb{R}^d$ is a finite-dimensional parameter space), an in-processing fair ML routine is to minimize the (empirical) risk $\mathbb{E}\big[\ell(\theta; X, Y)\big]$ while satisfying some fairness constraints. To keep things simple, we assume there are only two demographic groups; *i.e.* $|\mathcal{A}| = 2$. Without loss of generality, we refer to one group as advantaged ($A = 1$) and the other as disadvantaged ($A = 0$).

Consider fair learning with demographic parity as an example. At the population level, the goal is to solve the problem:

$$\theta^\star \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \mathbb{E}\big[\ell(\theta; X, Y)\big] \\ \text{subject to} & \mathbb{E}\big[\mathbf{1}\{f_\theta(X) = 1\}|A = 1\big] - \mathbb{E}\big[\mathbf{1}\{f_\theta(X) = 1\}|A = 0\big] = 0 \end{array} \right\}, \tag{3.2.1}$$

where the expectation is with respect to the distribution of tuple (X, A, Y). The true underlying distribution is unknown, so we cannot solve (3.2.1) directly. Instead, we observe IID training samples $\{(X_i, A_i, Y_i)\}_{i=1}^n$ from the true distribution and solve the empirical version of (3.2.1):

$$\widehat{\theta}_n \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \frac{1}{n}\sum_{i=1}^n \ell(\theta; X_i, Y_i) \\ \text{subject to} & \left| \frac{\sum_{i=1}^n \mathbf{1}\{f_\theta(X_i)=1, A_i=1\}}{\sum_{i=1}^n \mathbf{1}\{A_i=1\}} - \frac{\sum_{i=1}^n \mathbf{1}\{f_\theta(X_i)=1, A_i=0\}}{\sum_{i=1}^n \mathbf{1}\{A_i=0\}} \right| \le \alpha_n \end{array} \right\}, \tag{3.2.2}$$

where $0 < \alpha_n = o(\frac{1}{\sqrt{n}})$ is a slackness term shrinking to zero at a rate faster than $\frac{1}{\sqrt{n}}$. Through the rest of the work, we always let α_n be a positive number of order $o(\frac{1}{\sqrt{n}})$. In addition, we mention that our theory (see Section 3.3) can be applied not only to classification tasks, but also to regression tasks and, more generally, to all tasks under the empirical risk minimization learning paradigm.

3.2.2 Fair Machine Learning with Noisy Sensitive Attributes

Access to the true sensitive attribute A can be restricted due to data privacy or data missingness. Instead, the learner have access to the noisy sensitive attribute Z which can be served as a proxy for the true A. One may wish to learn a fair classifier by directly enforcing fairness notion on Z_i's, the proxies for A_i's. This approach is feasible and justifiable if the proxy constraint function is an unbiased estimate of the true constraint function modulo a constant multiplier.

Now we consider fair ML with noisy sensitive attributes by (empirical) risk minimization subject to fairness constraints with respect to Z. Take fair learning with demographic parity as an example. At the population level, the goal is to solve the problem

$$\theta^\star \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \mathbb{E}\big[\ell(\theta; X, Y)\big] \\ \text{subject to} & \mathbb{E}\big[\mathbf{1}\{f_\theta(X)=1\}|Z=1\big] - \mathbb{E}\big[\mathbf{1}\{f_\theta(X)=1\}|Z=0\big] = 0 \end{array} \right\}, \tag{3.2.3}$$

where the expectation is with respect to the distribution of tuple (X, Z, Y). Here Z is the proxy sensitive attribute but the true sensitive attribute A is unobservable. The true underlying distribution is unknown, so we cannot solve (3.2.3) directly. Instead, we observe IID (private) training samples $\{(X_i, Z_i, Y_i)\}_{i=1}^n$ from the true distribution and solve the empirical version of (3.2.3):

$$\widetilde{\theta}_n \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \frac{1}{n}\sum_{i=1}^n \ell(\theta; X_i, Y_i) \\ \text{subject to} & \left| \frac{\sum_{i=1}^n \mathbf{1}\{f_\theta(X_i)=1, Z_i=1\}}{\sum_{i=1}^n \mathbf{1}\{Z_i=1\}} - \frac{\sum_{i=1}^n \mathbf{1}\{f_\theta(X_i)=1, Z_i=0\}}{\sum_{i=1}^n \mathbf{1}\{Z_i=0\}} \right| \le \alpha_n \end{array} \right\}. \tag{3.2.4}$$

Let $c(\theta)$ and $\widetilde{c}(\theta)$ be the true constraint function in (3.2.1) and the proxy constraint function in (3.2.3) respectively. If $\widetilde{c}(\theta)$ equals $c(\theta)$ for all $\theta \in \Theta$ up to a constant multiplier, that is, $\widetilde{c}(\theta) \equiv \psi c(\theta)$ for some constant $\psi \in \mathbb{R}$, then the population problem (3.2.1) and (3.2.3) have exactly the same solution $\theta^\star$ (assuming uniqueness

of the solution). One can also show that under regularity conditions both $\widehat{\theta}_n$ and $\widetilde{\theta}_n$, the solution to (3.2.2) and to (3.2.4), are $\sqrt{n}$-consistent for $\theta^\star$. Our goal is to compare the estimating quality of $\widehat{\theta}_n$ and $\widetilde{\theta}_n$, and quantify the quality difference in terms of problem-specific parameters, for example, privacy level in a differential privacy mechanism (see Section 3.3.1), and missing proportion in a missing data mechanism (see Section 3.3.2).

3.2.3 Asymptotic Relative Efficiency

In statistics, *consistency* and *efficiency* are popular notions to evaluate the performance of estimators.

Definition 3.2.3 (Consistency)**.** An estimator $\widehat{\theta}_n$ is consistent for $\theta^\star$ if $\widehat{\theta}_n \xrightarrow{p} \theta^\star$ as $n \to \infty$.

Suppose that we have two consistent estimators $\widehat{\theta}_n$ and $\widetilde{\theta}_n$. Both of them are reasonable, but which one should be preferred? To answer this question, we can employ the notion of efficiency, *i.e.* measuring how spread out about $\widehat{\theta}_n$ (or $\widetilde{\theta}_n$) is the sampling distribution of the estimator. In light of this, we now adapt the concept of statistical efficiency to fair machine learning.

In fair ML, the most important metric to evaluate the performance of a classifier is fairness violation. Let $c : \Theta \to \mathbb{R}$ be the constraint function. For example, demographic parity constraint corresponds to $c(\theta) = \mathbb{E}\big[\mathbf{1}\{f_\theta(X) = 1\}|A = 1\big] - \mathbb{E}\big[\mathbf{1}\{f_\theta(X) = 1\}|A = 0\big]$. Since the exact fairness notion entails a classifier f_θ is fair if $c(\theta) = 0$, we define the *(signed) fairness violation* of θ as $c(\theta)$ itself.

Definition 3.2.4 (Efficiency in terms of constraint violations)**.** Suppose that we have two consistent estimators $\widehat{\theta}_n$ and $\widetilde{\theta}_n$ satisfying

$$\sqrt{n}\{c(\widehat{\theta}_n) - \cancel{c(\theta^\star)}\} \xrightarrow{d} \mathcal{N}(0, \sigma^2) \text{ and } \sqrt{n}\{c(\widetilde{\theta}_n) - \cancel{c(\theta^\star)}\} \xrightarrow{d} \mathcal{N}(0, \widetilde{\sigma}^2)$$

as $n \to \infty$. We say that the estimator $\widehat{\theta}_n$ is more efficient (in terms of constraint violations) than $\widetilde{\theta}_n$ if $\sigma^2 \leq \widetilde{\sigma}^2$. The asymptotic relative efficiency (ARE) of $\widetilde{\theta}_n$ to $\widehat{\theta}_n$ is

$$\mathrm{ARE}(\widetilde{\theta}_n, \widehat{\theta}_n) \triangleq \frac{\sigma^2}{\widetilde{\sigma}^2}.$$

In other words, the estimator $\widehat{\theta}_n$ is more efficient than $\widetilde{\theta}_n$ if $\mathrm{ARE}(\widetilde{\theta}_n, \widehat{\theta}_n) \leq 1$.

The first-order Taylor approximation to $c(\theta)$ around $\theta^\star$ is given by

$$c(\widehat{\theta}) = \cancel{c(\theta^\star)} + \nabla c(\theta^\star)^\top (\widehat{\theta} - \theta^\star) + o(\|\widehat{\theta} - \theta^\star\|_2), \tag{3.2.5}$$

where $\nabla c(\theta^\star)$ is the the gradient of $c(\cdot)$ evaluated at $\theta^\star$. From (3.2.5) we see that the constraint violations only depend on the part of the estimation error $\widehat{\theta} - \theta^\star$ in $\mathrm{ran}(\nabla c(\theta^\star))$, the range space of the linear transform $x \mapsto \nabla c(\theta^\star)^\top x : \mathbb{R}^d \supset \Theta \to \mathbb{R}$. Therefore, the efficiency in terms of constraint violations only concerns how fast an estimator converges to $\theta^\star$ in $\mathrm{ran}(\nabla c(\theta^\star))$.

Another way to examine the efficiency loss is to look at the asymptotic joint distribution of $c(\widehat{\theta}_n)$ and $c(\widetilde{\theta}_n)$. Let ρ be the asymptotic correlation between $c(\widehat{\theta}_n)$ and $c(\widetilde{\theta}_n)$. The fairness violations of the two estimators can be compared using the ratio of $c(\widehat{\theta}_n)$ to $c(\widetilde{\theta}_n)$, which converges in distribution to a Cauchy random variable U:

$$\frac{c(\widehat{\theta}_n)}{c(\widetilde{\theta}_n)} \xrightarrow{d} U \sim p_U(u) = \frac{1}{\pi} \frac{\beta}{(u-\alpha)^2 + \beta^2} \text{ with } \alpha = \frac{\rho\sigma}{\widetilde{\sigma}}, \beta = \frac{\sigma}{\widetilde{\sigma}}\sqrt{1-\rho^2}.$$

Constraint violation *inflates* if we observe a value of the ratio $|c(\widehat{\theta}_n)/c(\widetilde{\theta}_n)|$ less than one. Assume $\widetilde{\theta}_n$ is more efficient than $\widehat{\theta}_n$, *i.e.* $\sigma^2 < \widetilde{\sigma}^2$. Since $|\rho| \leq 1$, the median and mode of U, α, satisfies $|\alpha| < 1$, which indicates a high likelihood of constraint violation inflation. Precisely, the asymptotic probability of constraint violation inflation

is

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{c(\widehat{\theta}_n)}{c(\widetilde{\theta}_n)}\right| < 1\right) = \frac{1}{\pi}\left\{\tan^{-1}\left(\frac{\widetilde{\sigma}-\rho\sigma}{\sigma\sqrt{1-\rho^2}}\right) + \tan^{-1}\left(\frac{\widetilde{\sigma}+\rho\sigma}{\sigma\sqrt{1-\rho^2}}\right)\right\} > \frac{1}{2}.$$

In the rest of the chapter, the asymptotic relative efficiency (ARE) is the key quantity of interest, which compares the asymptotic variances of two estimators by $\mathrm{ARE} = \lim_{n\to\infty} \mathrm{Var}[c(\widehat{\theta}_n)]/\mathrm{Var}[c(\widetilde{\theta}_n)]$. We also empirically check the probability of constraint violation inflation in real data experiments (see Section 3.5).

3.3 Noise Cost in Fair Machine Learning

In this section, we wish to study $\mathrm{ARE}(\widetilde{\theta}_n, \widehat{\theta}_n)$, the asymptotic relative efficiency (ARE) of $\widetilde{\theta}_n$ to $\widehat{\theta}_n$ given by solving (3.2.4) and (3.2.2). To this end, we extend the notion of demographic parity and equality of opportunity to a more general form: we say that θ is fair (with respect to A) if

$$c(\theta) \triangleq \frac{\mathbb{E}\big[g(\theta; X, Y)|A=1\big]}{\mathbb{E}\big[h(X,Y)|A=1\big]} - \frac{\mathbb{E}\big[g(\theta; X, Y)|A=0\big]}{\mathbb{E}\big[h(X,Y)|A=0\big]} = 0. \tag{3.3.1}$$

The fairness notion (3.3.1) is known as *linear-fractional* fairness constraint (Celis et al., 2021), which encompasses a broad class of group fairness notions. To name a few, we note that demographic parity is a special case of (3.3.1) if we take $g(\theta; X, Y) = \mathbf{1}\{f_\theta(X) = 1\}$ and $h \equiv 1$. Besides, (3.3.1) becomes equality of opportunity if we take $g(\theta; X, Y) = \mathbf{1}\{f_\theta(X) = 1, Y = 1\}$ and $h(X, Y) = \mathbf{1}\{Y = 1\}$. When $h \equiv 1$, (3.3.1) degenerates to *linear* fairness (see discussion in Appendix B.1).

Let the marginal distribution of A and conditional distribution of (X, Y) given A be

$$\begin{cases} \mathbb{P}(A=0) = \pi_0, & \mathbb{P}(A=1) = \pi_1 \\ (X,Y)|A=0 \sim Q_0, & (X,Y)|A=1 \sim Q_1 \end{cases}. \tag{3.3.2}$$

Then the distribution of (X, A, Y) is uniquely identified by (3.3.2). Moreover, $(X, Y) \sim \pi_0 Q_0 + \pi_1 Q_1$ is a mixture of Q_0 and Q_1 weighted by π_0 and π_1. Also, denote the marginal distribution of Z and conditional distribution of (X, Y) given Z by

$$\begin{cases} \mathbb{P}(Z=0) = \widetilde{\pi}_0, & \mathbb{P}(Z=1) = \widetilde{\pi}_1 \\ (X,Y)|Z=0 \sim \widetilde{Q}_0, & (X,Y)|Z=1 \sim \widetilde{Q}_1 \end{cases}.$$

Enforcing fairness notion (3.3.1) with respect to Z is

$$\widetilde{c}(\theta) \triangleq \frac{\mathbb{E}\big[g(\theta; X, Y)|Z=1\big]}{\mathbb{E}\big[h(X,Y)|Z=1\big]} - \frac{\mathbb{E}\big[g(\theta; X, Y)|Z=0\big]}{\mathbb{E}\big[h(X,Y)|Z=0\big]} = 0.$$

Now we are almost ready to show the privacy cost in linear-fractional fairness (3.3.1)-aware learning. First, let the true parameter $\theta^\star$, *i.e.* the solution to the population problem, be

$$\theta^\star \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \mathbb{E}\big[\ell(\theta; X, Y)\big] \\ \text{subject to} & \dfrac{\mathbb{E}\big[g(\theta; X, Y)|A=1\big]}{\mathbb{E}\big[h(X,Y)|A=1\big]} - \dfrac{\mathbb{E}\big[g(\theta; X, Y)|A=0\big]}{\mathbb{E}\big[h(X,Y)|A=0\big]} = 0 \end{array} \right\}, \tag{3.3.3}$$

where the expectation is with respect to the underlying distribution of tuple (X, A, Y).

Then, let the estimator $\widehat{\theta}_n$ be the solution to the empirical problem given the true sensitive attribute,

$$\widehat{\theta}_n \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \frac{1}{n}\sum_{i=1}^n \ell(\theta; X_i, Y_i) \\ \text{subject to} & \left| \frac{\sum_{i=1}^n g(\theta; X_i, Y_i)\mathbf{1}\{A_i=1\}}{\sum_{i=1}^n h(X_i, Y_i)\mathbf{1}\{A_i=1\}} - \frac{\sum_{i=1}^n g(\theta; X_i, Y_i)\mathbf{1}\{A_i=0\}}{\sum_{i=1}^n h(X_i, Y_i)\mathbf{1}\{A_i=0\}} \right| \le \alpha_n \end{array} \right\}.$$

Finally, let $\widetilde{\theta}_n$ be the solution to the empirical problem given the proxy sensitive

attribute,

$$\widetilde{\theta}_n \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \frac{1}{n}\sum_{i=1}^n \ell(\theta; X_i, Y_i) \\ \text{subject to} & \left| \frac{\sum_{i=1}^n g(\theta;X_i,Y_i)\mathbf{1}\{Z_i=1\}}{\sum_{i=1}^n h(X_i,Y_i)\mathbf{1}\{Z_i=1\}} - \frac{\sum_{i=1}^n g(\theta;X_i,Y_i)\mathbf{1}\{Z_i=0\}}{\sum_{i=1}^n h(X_i,Y_i)\mathbf{1}\{Z_i=0\}} \right| \leq \alpha_n \end{array} \right\}. \quad (3.3.4)$$

We made the following technical assumptions on the population problem (3.3.3).

1. **smoothness and concentration:** ℓ and g are twice continuously differentiable with respect to θ, and $\ell(\theta^\star; X, Y)$, $\nabla\ell(\theta^\star; X, Y)$, $g(\theta^\star; X, Y)$, $\nabla g(\theta^\star; X, Y)$, $h(X, Y)$ are sub-Gaussian.

2. **uniqueness:** the stochastic optimization problem with a single expected value constraint (3.3.3) has unique optimal primal-dual pair $(\theta^\star, \lambda^\star)$, and $\theta^\star$ belongs to the interior of the compact set Θ.

3. **positive definiteness:** The Hessian of the Lagrangian evaluated at $(\theta^\star, \lambda^\star)$ is positive definite.

The preceding assumptions are not the most general, but they are easy to interpret. The smoothness conditions on ℓ and g with respect to θ, the concentration conditions of $\ell(\theta^\star)$, $g(\theta^\star)$ and h, and the uniqueness condition facilitate the use of standard tools from asymptotic statistics to study the large sample properties of the constraint value. The positive definiteness condition postulates the Lagrangian of the equality constrained optimization problem is locally strongly convex at $(\theta^\star, \lambda^\star)$.

The main technical result characterizes the efficiency of $\widehat{\theta}_n$ and $\widetilde{\theta}_n$ (see proof in Appendix B.3).

Theorem 3.3.1 (Noise cost in linear-fractional fairness (3.3.1)-aware learning)**.** *Suppose that $\widetilde{c}(\theta) \equiv \psi c(\theta)$ for some constant $\psi \in \mathbb{R}$. Under the standing assumptions, let estimators $\widehat{\theta}_n$ and $\widetilde{\theta}_n$ be consistent for $\theta^\star$, then*

$$\sqrt{n}\{c(\widehat{\theta}_n) - \cancel{c(\theta^\star)}\} \xrightarrow{d} \mathcal{N}(0, \sigma^2) \text{ and } \sqrt{n}\{c(\widetilde{\theta}_n) - \cancel{c(\theta^\star)}\} \xrightarrow{d} \mathcal{N}(0, \widetilde{\sigma}^2),$$

where

$$\sigma^2 = \frac{\mathrm{Var}_{Q_0}[g(\theta^\star; X, Y) - \kappa h(X, Y)]}{\pi_0(\mathbb{E}_{Q_0}[h(X, Y)])^2} + \frac{\mathrm{Var}_{Q_1}[g(\theta^\star; X, Y) - \kappa h(X, Y)]}{\pi_1(\mathbb{E}_{Q_1}[h(X, Y)])^2},$$

$$\widetilde{\sigma}^2 = \psi^{-2} \times \left\{ \frac{\mathrm{Var}_{\widetilde{Q}_0}[g(\theta^\star; X, Y) - \kappa h(X, Y)]}{\widetilde{\pi}_0(\mathbb{E}_{\widetilde{Q}_0}[h(X, Y)])^2} + \frac{\mathrm{Var}_{\widetilde{Q}_1}[g(\theta^\star; X, Y) - \kappa h(X, Y)]}{\widetilde{\pi}_1(\mathbb{E}_{\widetilde{Q}_1}[h(X, Y)])^2} \right\},$$

and

$$\kappa \triangleq \frac{\mathbb{E}_{Q_0}[g(\theta^\star; X, Y)]}{\mathbb{E}_{Q_0}[h(X, Y)]} = \frac{\mathbb{E}_{Q_1}[g(\theta^\star; X, Y)]}{\mathbb{E}_{Q_1}[h(X, Y)]}.$$

Theorem 3.3.2 offers a general recipe for the evaluation of asymptotic relative efficiency (in terms of constraint violation) of $\widetilde{\theta}_n$ to $\widehat{\theta}_n$. With additional distributional structure on the noisy sensitive attribute Z, we can examine the effect of problem-dependent parameters (*e.g.*, privacy level and missing proportion) on $\mathrm{ARE}(\widetilde{\theta}_n, \widehat{\theta}_n) = \sigma^2/\widetilde{\sigma}^2$.

3.3.1 Application 1: Fair ML with Private Sensitive Attributes

Consider the randomized response mechanism (Warner, 1965; Kairouz et al., 2014) for releasing privatized sensitive attribute:

$$Q(Z = z \mid A = a) = \begin{cases} \frac{e^\varepsilon}{|\mathcal{A}|-1+e^\varepsilon} & \text{if } z = a \\ \frac{1}{|\mathcal{A}|-1+e^\varepsilon} & \text{if } z \neq a \end{cases} \tag{3.3.5}$$

for all $a, z \in \mathcal{A}$, where $\varepsilon > 0$ controls the privacy level. The *privatized sensitive attribute* Z of the true sensitive attribute A is defined as $Z = Q(\cdot \mid A)$. In addition, the sampling mechanism Q requires $Z \perp\!\!\!\perp X, Y \mid A$. Then the private dataset $\{(X_i, Z_i, Y_i)\}_{i=1}^n$ is generated from the original dataset $\{(X_i, A_i, Y_i)\}_{i=1}^n$ via the transition kernel Q.

The randomized response mechanism (3.3.5) is a locally ε-differentially private

mechanism (Duchi et al., 2013), that is

$$\max_{z,a,a' \in \mathcal{A}} \frac{Q(Z = z \mid A = a)}{Q(Z = z \mid A = a')} \leq e^{\varepsilon}.$$

Here a smaller parameter ε indicates a stronger privacy guarantee. Moreover, the mechanism Q is considered optimal for distribution estimation under local differential privacy constraints (Kairouz et al., 2014, 2016).

From this point forward (with the exception of the general theory presented in Section 3.3.3), we assume that there are only two demographic groups, *i.e.* $|\mathcal{A}| = 2$. The mechanism (3.3.5) becomes

$$Q(Z = z \mid A = a) = \begin{cases} \frac{e^{\varepsilon}}{1+e^{\varepsilon}} \triangleq 1 - \gamma & \text{if } z = a \\ \frac{1}{1+e^{\varepsilon}} \triangleq \gamma & \text{if } z \neq a \end{cases} \tag{3.3.6}$$

for $a \in \{0, 1\}$, where $\gamma \in [0, 0.5)$. The parameter $\gamma = 0$ (or equivalently $\varepsilon = \infty$) signifies complete lack of privacy, whereas $\gamma \to 0.5$ (or equivalently $\varepsilon \to 0$) corresponds to perfect privacy.

Proposition 3.3.1 (Invariance of exact linear-fractional fairness under local differential privacy)**.** *Consider any exact fairness notion defined by linear-fractional fairness* (3.3.1)*. Let* $\widehat{Y} = f(X)$ *be a binary classifier. Under the local differential privacy mechanism* (3.3.6)*,* $\widehat{Y}$ *is fair with respect to* A *if and only if* $\widehat{Y}$ *is fair with respect to* Z*.*

Proof. By some algebra, we have that the proxy constraint function $\widetilde{c}(\theta)$ is equal to the true constraint function $c(\theta)$ up to a scaling factor: $\widetilde{c}(\theta) = \psi_{\text{frac}}(\gamma, \pi_0, \pi_1, m_0, m_1) \times c(\theta)$, where

$$\psi_{\text{frac}}(\gamma, \pi_0, \pi_1, m_0, m_1) \triangleq \frac{(1 - 2\gamma)\pi_0\pi_1 m_0 m_1}{\{\gamma\pi_0 m_0 + (1-\gamma)\pi_1 m_1\}\{(1-\gamma)\pi_0 m_0 + \gamma\pi_1 m_1\}},$$

as well $m_0 \triangleq \mathbb{E}_{Q_0}\big[h(X,Y)\big]$ and $m_1 \triangleq \mathbb{E}_{Q_1}\big[h(X,Y)\big]$. This implies $c(\theta) = 0$ if and only if $\widetilde{c}(\theta) = 0$. □

Proposition 3.3.1 requires $\widehat{Y}$ is only a function of X. Mozannar et al. (2020) show by construction the existence of a classifier $\widehat{Y} = \widetilde{f}(X, Z)$ which is fair in terms of demographic parity with respect to Z but unfair to A.

Theorem 3.3.2 (Privacy cost in linear-fractional fairness (3.3.1)-aware learning)**.** *Under the standing assumptions, let estimators $\widehat{\theta}_n$ and $\widetilde{\theta}_n$ be consistent for $\theta^\star$. The asymptotic relative efficiency (ARE) of $\widetilde{\theta}_n$ to $\widehat{\theta}_n$ is*

$$\mathrm{ARE}(\widetilde{\theta}_n, \widehat{\theta}_n) = \varphi\left(\gamma, \frac{\pi_0 m_0}{\pi_1 m_1}, \frac{\mathrm{Var}[g(\theta^\star; X, Y) - \kappa h(X,Y) | A = 0]/m_0}{\mathrm{Var}[g(\theta^\star; X, Y) - \kappa h(X,Y) | A = 1]/m_1}\right), \tag{3.3.7}$$

where

$$\varphi(\gamma, r_1, r_2) \triangleq \frac{(1-2\gamma)^2 r_1 (r_1 + r_2)}{\{\gamma r_1 + (1-\gamma)\}^2\{(1-\gamma) r_1 r_2 + \gamma\} + \{(1-\gamma) r_1 + \gamma\}^2\{\gamma r_1 r_2 + (1-\gamma)\}}.$$

Recall that demographic parity corresponds to $h \equiv 1$ and equality of opportunity corresponds to $h(X,Y) = \mathbf{1}\{Y = 1\}$. In order to interpret (3.3.7), we therefore take $h(X,Y) = \mathbf{1}\{\mathcal{E}(X,Y)\}$, where $\mathcal{E}(X,Y)$ is an event of X and Y. Then the ARE (3.3.7) becomes

$$\mathrm{ARE}(\widetilde{\theta}_n, \widehat{\theta}_n) = \varphi\left(\gamma, \frac{\mathbb{P}(\mathcal{E}(X,Y), A = 0)}{\mathbb{P}(\mathcal{E}(X,Y), A = 1)}, \frac{\mathrm{Var}[g(\theta^\star; X, Y) | \mathcal{E}(X,Y), A = 0]}{\mathrm{Var}[g(\theta^\star; X, Y) | \mathcal{E}(X,Y), A = 1]}\right).$$

Note that the ARE is jointly determined by the level of privacy, a ratio of marginal probabilities of the minority and majority groups, and a ratio of their conditional variances. Theorem 3.3.2 demonstrates that the cost of privacy is the efficiency loss in terms of fairness violations. For fixed ratios

$$r_1 \triangleq \frac{\mathbb{P}(\mathcal{E}(X,Y), A = 0)}{\mathbb{P}(\mathcal{E}(X,Y), A = 1)} > 0,$$

and

$$r_2 \triangleq \frac{\mathrm{Var}[g(\theta^\star; X, Y)|\mathcal{E}(X, Y), A = 0]}{\mathrm{Var}[g(\theta^\star; X, Y)|\mathcal{E}(X, Y), A = 1]} > 0,$$

function $\varphi(\gamma, r_1, r_2)$ is decreasing in γ. In the absence of privacy, $\varphi(0, r_1, r_2) = 1$ means no efficiency loss. Under perfect privacy, $\varphi(0.5, r_1, r_2) = 0$ indicates total loss of efficiency. Moreover, $\widehat{\theta}_n$ is always more efficient than $\widetilde{\theta}_n$ because $\mathrm{ARE}(\widetilde{\theta}_n, \widehat{\theta}_n) \leq 1$.

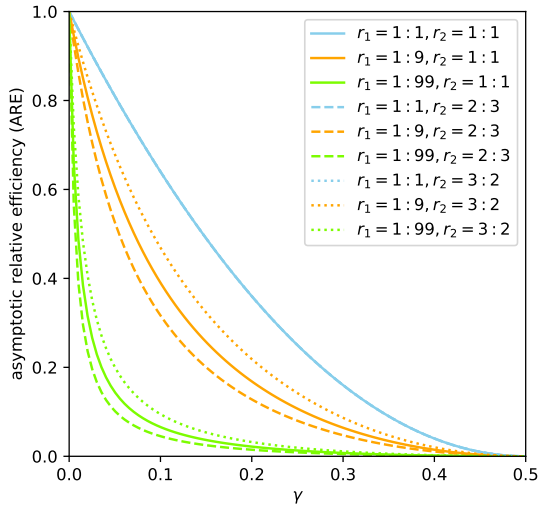

Figure III.1: Asymptotic relative efficiency curve of γ for varying r_1 and r_2.

Figure III.1 demonstrates the asymptotic relative efficiency (ARE) curve of privacy level γ for varying ratios r_1 and r_2. The ARE is always upper bounded by $(1 - 2\gamma)^2$, which is achieved only if $r_1 = 1$. Therefore for any fixed γ and r_2, the ARE achieves its maximum only if the dataset is balanced in the sense that $\mathbb{P}(\mathcal{E}(X, Y), A = 0) = \mathbb{P}(\mathcal{E}(X, Y), A = 1)$. Moreover, for any fixed γ and r_2, the ARE is strictly increasing in r_1 (assuming $r_1 \leq 1$). This implies the effect of subpopulation size imbalance: demographic group imbalance degrades the efficiency loss in privately fair learning. In the literature, the effect of group size imbalance on the difficulty of learning fair classifier from contaminated data (note that private sensitive attribute is a particular type of data contamination) was also reported in Konstantinov and Lampert (2022) and the references therein. Lastly, the ARE is strictly increasing in the problem-specific parameter r_2, given fixed γ and $r_1 < 1$.

3.3.2 Application 2: Fair ML with Missing Sensitive Attributes

Some users may choose not to disclose their demographic identities during data collection due to privacy concerns. We investigate how the absence of sensitive attributes impacts the generalizability of fairness constraints. Consider the following missing data mechanism for sensitive attributes :

$$\mathbb{P}(R = 1 \mid X, A, Y) = \mathbb{P}(R = 1 \mid A) \triangleq \omega_A. \tag{3.3.8}$$

where $R = 1$ corresponds to response (*i.e.*, A is observed) and otherwise $R = 0$ corresponds to non-response (*i.e.*, A is missing). The missingness mechanism (3.3.8) is a particular type of missing at random (MAR) at the population level and missing completely at random (MCAR) within each subpopulation. By introducing a dummy group -1, we can encode the noisy sensitive attribute Z by

$$Z = \begin{cases} A & \text{if } R = 1 \\ -1 & \text{if } R = 0 \end{cases}, \tag{3.3.9}$$

which means that missing values are all annotated as -1. This corresponds to the transition kernel

$$Q(Z = z \mid A = a) = \begin{cases} \omega_a & \text{if } z = a \\ 1 - \omega_a & \text{if } z = -1 \end{cases}. \tag{3.3.10}$$

One common approach for analyzing data with missing values is to just use the completely observed samples (i.e., samples with all features observed) and discard the samples with some missing features. We employ this strategy by solving the empirical problem (3.3.4), of which the empirical risk function is computed with all samples while the fairness constraint function is calculated with samples that include

the sensitive attribute.

From now on, we assume that the response probability is non-vanishing, *i.e.*, $\omega_a > 0$ for $a \in \{0, 1\}$.

Proposition 3.3.2 (Invariance of exact linear-fractional fairness under missing at random)**.** *Consider any exact fairness notion defined by linear-fractional fairness* (3.3.1)*. Let* $\widehat{Y} = f(X)$ *be a binary classifier. Under the missing data mechanism* (3.3.8)*,* $\widehat{Y}$ *is fair with respect to* A *if and only if* $\widehat{Y}$ *is fair with respect to* Z*.*

Proof. By the MAR mechanism, we have that the proxy constraint function $\widetilde{c}(\theta)$ is equal to the true constraint function $c(\theta)$: $\widetilde{c}(\theta) = c(\theta)$. This implies $c(\theta) = 0$ if and only if $\widetilde{c}(\theta) = 0$. □

Theorem 3.3.3 (Missingness cost in linear-fractional fairness (3.3.1)-aware learning)**.** *Under the standing assumptions, let estimators* $\widehat{\theta}_n$ *and* $\widetilde{\theta}_n$ *be consistent for* $\theta^\star$*. The asymptotic relative efficiency (ARE) of* $\widetilde{\theta}_n$ *to* $\widehat{\theta}_n$ *is*

$$\mathrm{ARE}(\widetilde{\theta}_n, \widehat{\theta}_n) = \frac{r_2 + r_1}{\omega_0^{-1} r_2 + \omega_1^{-1} r_1}, \tag{3.3.11}$$

where

$$r_1 = \frac{\pi_0 m_0}{\pi_1 m_1} \quad \textit{and} \quad r_2 = \frac{\mathrm{Var}[g(\theta^\star; X, Y) - \kappa h(X, Y) | A = 0]/m_0}{\mathrm{Var}[g(\theta^\star; X, Y) - \kappa h(X, Y) | A = 1]/m_1}.$$

Theorem 3.3.3 indicates that any probability of missing data degrades the asymptotic efficiency of the estimator inversely proportionally. Consider a working model $\omega_1 = 1 - p$ and $\omega_0 = (1 - p)^\alpha$, where $p \in (0, 1)$ controls the missing proportion and $\alpha > 0$ controls the functional relationship between ω_1 and ω_0. In particular, group 1 has p missing proportion while group 0 has $1 - (1 - p)^\alpha$. Figure III.2 demonstrates the ARE curves of p with $\alpha \in \{0.5, 1, 2\}$, which is strictly decreasing in the missing proportion p given fixed α.

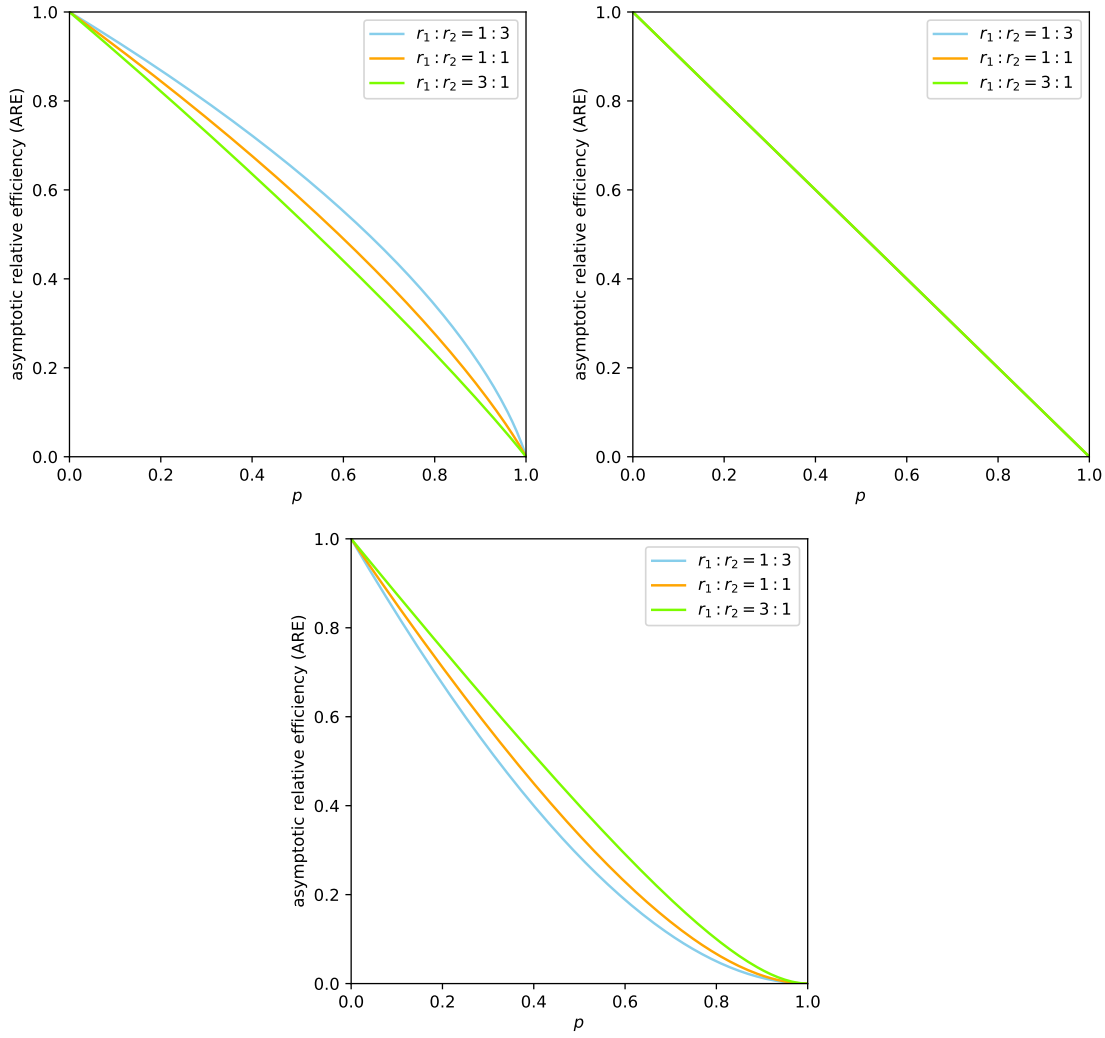

Figure III.2: Asymptotic relative efficiency curve of p for $\alpha = 0.5$ (top left), $\alpha = 1$ (top right), and $\alpha = 2$ (bottom).

3.3.3 General Theory

In this subsection, we extend our theory of two demographic groups to general number of groups. Suppose we have $K+1 (K \geq 2)$ groups indexed by $0, 1, \ldots, K$. The notion of linear-fractional fairness (3.3.1) can be adapted to more than two groups: we say θ is fair if

$$\frac{\mathbb{E}\big[g(\theta; X, Y)|A = k\big]}{\mathbb{E}\big[h(X, Y)|A = k\big]} - \frac{\mathbb{E}\big[g(\theta; X, Y)|A = 0\big]}{\mathbb{E}\big[h(X, Y)|A = 0\big]} = 0 \quad \text{for } k \in [K], \tag{3.3.12}$$

where group 0 is referred to as a reference group. One can also define a fairness notion by

$$\frac{\mathbb{E}\big[g(\theta;X,Y)|A=k\big]}{\mathbb{E}\big[h(X,Y)|A=k\big]}-\frac{\mathbb{E}\big[g(\theta;X,Y)\big]}{\mathbb{E}\big[h(X,Y)\big]}=0 \quad \text{for } k\in\{0\}\cup[K] \tag{3.3.13}$$

which is symmetric in group indices. Due to the equivalence of (3.3.13) and (3.3.12), we opt to use (3.3.12) for a comparison with two-group theory.

Let the marginal distribution of A and conditional distribution of (X,Y) given A be

$$\mathbb{P}(A=k)=\pi_k, (X,Y)|A=k\sim Q_k \quad \text{for } k\in\{0\}\cup[K]. \tag{3.3.14}$$

Then the distribution of (X,A,Y) is uniquely identified by (3.3.14). Moreover, the distribution of $(X,Y)\sim\sum_{k=0}^{K}\pi_k Q_k \overset{d}{=} Q_\star$ is a mixture of Q_k's weighted by π_k's.

The joint distribution of (X,Z,Y) is uniquely identified by the marginal distribution and conditional distribution as follows:

$$\mathbb{P}(A=k)=\widetilde{\pi}_k, (X,Y)|A=k\sim \widetilde{Q}_k \quad \text{for } k\in\{0\}\cup[K]. \tag{3.3.15}$$

Let the true parameter $\theta^\star$, *i.e.* the solution to the population problem, be

$$\theta^\star\in\begin{cases}\arg\min_{\theta\in\Theta} & \mathbb{E}\big[\ell(\theta;X,Y)\big]\\ \text{subject to} & \left\{\dfrac{\mathbb{E}\big[g(\theta;X,Y)|A=k\big]}{\mathbb{E}\big[h(X,Y)|A=k\big]}-\dfrac{\mathbb{E}\big[g(\theta;X,Y)|A=0\big]}{\mathbb{E}\big[h(X,Y)|A=0\big]}=0\right\}_{k=1}^{K}\end{cases},$$

where the expectation is with respect to the underlying distribution of tuple (X,A,Y).

Then, let the estimator $\widehat{\theta}_n$ be the solution to the empirical problem given the true sensitive attribute,

$$\widehat{\theta}_n\in\begin{cases}\arg\min_{\theta\in\Theta} & \frac{1}{n}\sum_{i=1}^{n}\ell(\theta;X_i,Y_i)\\ \text{subject to} & \left\{\left|\dfrac{\sum_{i=1}^{n}g(\theta;X_i,Y_i)\mathbf{1}\{A_i=k\}}{\sum_{i=1}^{n}h(X_i,Y_i)\mathbf{1}\{A_i=k\}}-\dfrac{\sum_{i=1}^{n}g(\theta;X_i,Y_i)\mathbf{1}\{A_i=0\}}{\sum_{i=1}^{n}h(X_i,Y_i)\mathbf{1}\{A_i=0\}}\right|\le\alpha_n\right\}_{k=1}^{K}\end{cases}.$$

Finally, let $\widetilde{\theta}_n$ be the solution to the empirical problem given the proxy sensitive attribute,

$$\widetilde{\theta}_n \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \frac{1}{n}\sum_{i=1}^n \ell(\theta; X_i, Y_i) \\ \text{subject to} & \left\{ \left| \frac{\sum_{i=1}^n g(\theta;X_i,Y_i)\mathbf{1}\{Z_i=k\}}{\sum_{i=1}^n h(X_i,Y_i)\mathbf{1}\{Z_i=k\}} - \frac{\sum_{i=1}^n g(\theta;X_i,Y_i)\mathbf{1}\{Z_i=0\}}{\sum_{i=1}^n h(X_i,Y_i)\mathbf{1}\{Z_i=0\}} \right| \leq \alpha_n \right\}_{k=1}^K \end{array} \right\}.$$

The true fairness constraint function $\boldsymbol{c}(\theta) : \mathbb{R}^d \to \mathbb{R}^K$ is defined as

$$\boldsymbol{c}(\theta) \triangleq (c_1(\theta), \ldots, c_K(\theta))^\top \text{ with } c_k(\theta) = \frac{\mathbb{E}\big[g(\theta; X, Y)|A = k\big]}{\mathbb{E}\big[h(X, Y)|A = k\big]} - \frac{\mathbb{E}\big[g(\theta; X, Y)|A = 0\big]}{\mathbb{E}\big[h(X, Y)|A = 0\big]}$$

for $k \in [K]$. Under the same assumptions as the two-group problem, we have the main technical result as follows (see Appendix B.4 for a complete treatment to the general-number-of-groups problem).

Theorem 3.3.4 (Noise cost in linear-fractional fairness (3.3.12)-aware learning)**.** *Suppose that $\widetilde{c}(\theta) \equiv \Psi c(\theta)$ for some constant matrix $\Psi \in \mathbb{R}^{K\times K}$. Under the standing assumptions, let estimators $\widehat{\theta}_n$ and $\widetilde{\theta}_n$ be consistent for $\theta^\star$, then*

$$\sqrt{n}\{\boldsymbol{c}(\widehat{\theta}_n) - \cancel{\boldsymbol{c}(\theta^\star)}\} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma) \text{ and } \sqrt{n}\{\boldsymbol{c}(\widetilde{\theta}_n) - \cancel{\boldsymbol{c}(\theta^\star)}\} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Psi^{-1}\widetilde{\Sigma}\Psi^{-\top}),$$

where

$$\Sigma_{kl} = \frac{\mathrm{Var}_{Q_0}[g(\theta^\star; X, Y) - \kappa h(X, Y)]}{\pi_0(\mathbb{E}_{Q_0}[h(X, Y)])^2} + \left(\frac{\mathrm{Var}_{Q_k}[g(\theta^\star; X, Y) - \kappa h(X, Y)]}{\pi_k(\mathbb{E}_{Q_k}[h(X, Y)])^2} \right) \mathbf{1}\{k = l\}$$

$$\widetilde{\Sigma}_{kl} = \frac{\mathrm{Var}_{\widetilde{Q}_0}[g(\theta^\star; X, Y) - \kappa h(X, Y)]}{\widetilde{\pi}_0(\mathbb{E}_{\widetilde{Q}_0}[h(X, Y)])^2} + \left(\frac{\mathrm{Var}_{\widetilde{Q}_k}[g(\theta^\star; X, Y) - \kappa h(X, Y)]}{\widetilde{\pi}_k(\mathbb{E}_{\widetilde{Q}_k}[h(X, Y)])^2} \right) \mathbf{1}\{k = l\}$$

for $k, l \in [K]$, and

$$\kappa \triangleq \frac{\mathbb{E}_{Q_0}[g(\theta^\star; X, Y)]}{\mathbb{E}_{Q_0}[h(X, Y)]} = \frac{\mathbb{E}_{Q_1}[g(\theta^\star; X, Y)]}{\mathbb{E}_{Q_1}[h(X, Y)]} = \ldots = \frac{\mathbb{E}_{Q_K}[g(\theta^\star; X, Y)]}{\mathbb{E}_{Q_K}[h(X, Y)]}.$$

3.4 Simulations

We simulate the asymptotic relative efficiency (ARE) for the risk-parity linear regression problem:

$$\begin{aligned} &\min_{\beta\in\Theta} \quad \mathbb{E}\big[(Y-\beta^\top X)^2\big] \\ &\text{subject to} \quad \mathbb{E}[(Y-\beta^\top X)^2|A=1]-\mathbb{E}[(Y-\beta^\top X)^2|A=0]=0 \end{aligned} \tag{3.4.1}$$

where we generate $n \in \{300, 3000\}$ samples by the following data generating process:

$$A \sim \text{Bernoulli}(1-\pi_0), X|A=a \sim \mathcal{N}(\mu_a, \Sigma_a) \text{ and } Y|X, A=a \sim \mathcal{N}(\beta_a^\top X, \sigma_a^2)$$

for $a \in \{0,1\}$. We pick $\mu_0 = (1,2)^\top, \mu_1 = (2,1)^\top, \Sigma_0 = \Sigma_1 = I_2, \sigma_0^2 = \sigma_1^2 = 1$ and investigate two scenarios: imbalanced subgroups with $\pi_0 = 0.3$ and balanced subgroups with $\pi_0 = 0.5$. The goal of the optimization problem (3.4.1) is to minimize the population risk (in least square) while satisfying the parity of subpopulation risks (in least square) of group $A=0$ and group $A=1$. The two-group risk-parity problem (3.4.1) can also be interpreted as a multi-objective optimization problem: to find the Pareto optimal point $\beta^\star$ of the bi-objective optimization problem $\min_{\beta\in\Theta}(\mathbb{E}[(Y-\beta^\top X)^2|A=1], \mathbb{E}[(Y-\beta^\top X)^2|A=0])$ such that $\mathbb{E}[(Y-\beta^{\star\top} X)^2|A=1] = \mathbb{E}[(Y-\beta^{\star\top} X)^2|A=0]$.

In Figure III.3, we plot relative efficiency curves for $\pi_0 = 0.3$ and $\pi_0 = 0.5$, all of which are averaged over 500 replicates. For large sample size n, the relative efficiency curves are close to the theoretical line of asymptotic relative efficiency curve, validating our theory in the large sample regime. As a by-product, our theory can visualize the fairness-privacy trade-off without retraining models with varying privacy budgets.

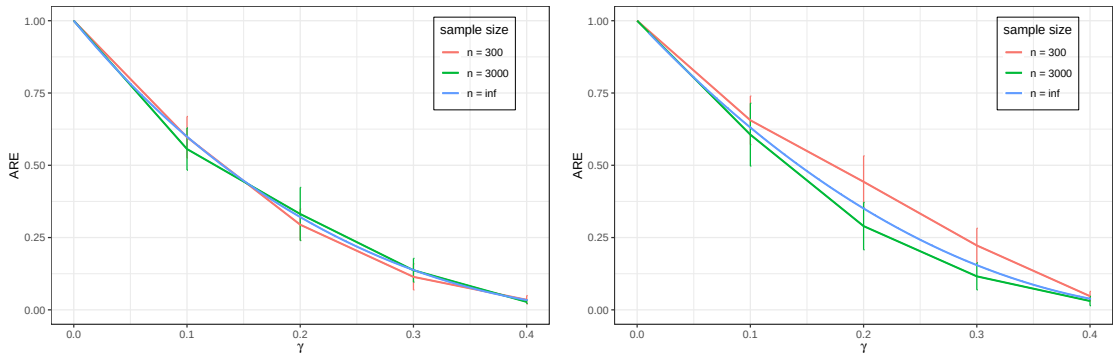

Figure III.3: Relative efficiency curves for $\pi_0 = 0.3$ (left) and $\pi_0 = 0.5$ (right).

3.5 Real Data Experiments

We conduct real data experiments using the `ACSIncome` dataset (Ding et al., 2021) derived from US Census surveys, which is a modern version of the `Adult` dataset (Dua and Graff, 2017) from UCI. The classification task is to predict whether an individual's income per year is higher than \$50K. The fairness goal is demographic parity (DP): $\mathbb{P}(\widehat{Y} = 1 \mid A = 1) - \mathbb{P}(\widehat{Y} = 1 \mid A = 0) = 0$, where $A = 1$ for white people is the advantaged group and $A = 0$ for non-white people is the disadvantaged group. We use a logistic regression model for classification and train a model by minimizing the logistic loss subject to the demographic parity constraint.

Experiments with private sensitive attributes. We release noisy sensitive attributes according to the local differential privacy mechanism (3.3.6) with $\gamma \in \{0, 0.1, 0.2, 0.3, 0.4\}$. Figure III.4 demonstrates boxplots and density plots for training and test signed constraint violation. The boxplots show that demographic parity is satisfied on training set but violated on test set, which indicates the issue of fairness generalizability for in-training fair learning approaches. Examine the density plots, as the privacy level γ increases, the distribution of test constraint violation is still centered around the origin but with heavier tails. Specifically, the change in spread of these distributions is characterized by the asymptotic relative efficiency (ARE) curve. The empirical curve coincides well with the theoretical curve as illustrated

by the bottom right panel of Figure III.4. Moreover, we report the frequency of constraint violation inflation in Table III.1 (see discussion in Section 3.2.3). The phenomenon of constraint violation inflation happens with high probability greater than 99% for 100 replicates.

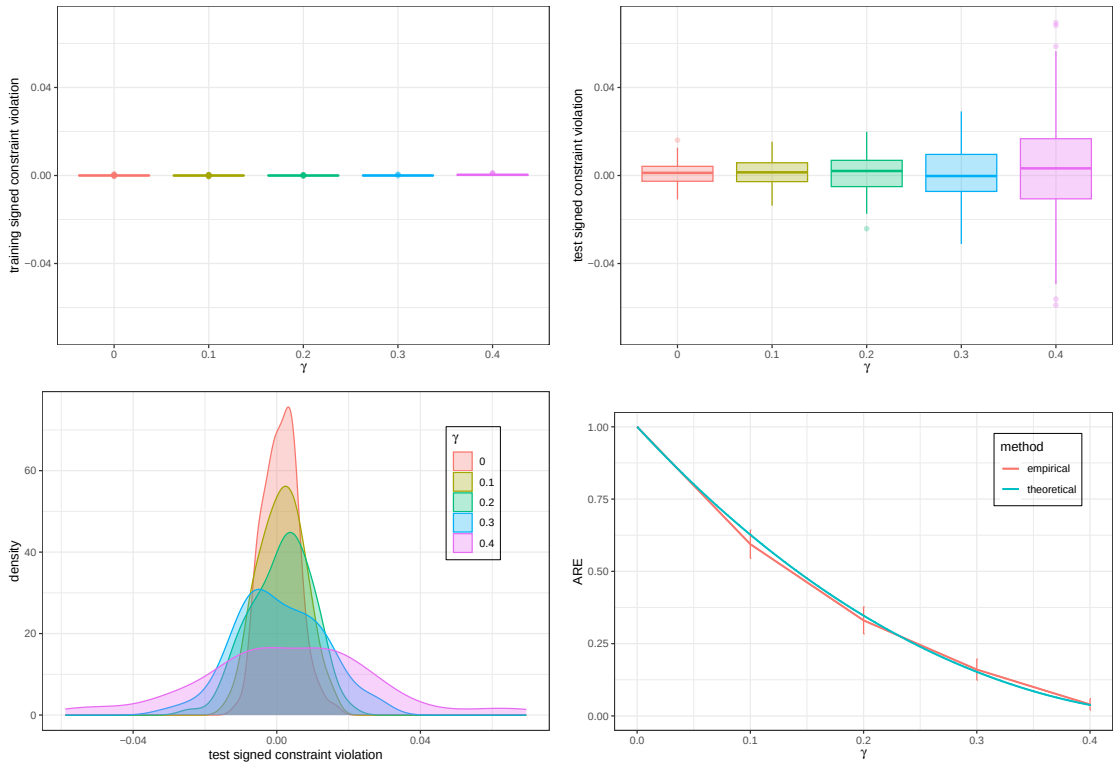

Figure III.4: Results of experiments with private sensitive attributes. Top left: boxplot of training signed constraint violation. Top right: boxplot of test signed constraint violation. Bottom left: density plot of test signed constraint violation. Bottom right: empirical and theoretical curves of relative efficiency.

γ	0	0.1	0.2	0.3	0.4
constraint violation inflation frequency	0.99	0.99	1.00	0.99	0.99

Table III.1: Results of experiments with private sensitive attributes. Frequency of constraint violation inflation.

Experiments with missing sensitive attributes. We release noisy sensitive attributes according to the missing data mechanism (3.3.8) with $\omega_0 = \omega_1 = p \in \{0, 0.1, 0.2, 0.3, 0.4\}$ so that the missing proportions of group 0 and group 1 are both p. Figure III.5 shows the empirical and theoretical curves of relative efficiency.

As suggested by Theorem 3.3.3, the asymptotic relative efficiency curve should be $\mathrm{ARE}(p) = 1 - p$. The empirical curve aligns well with the theoretical curve as shown by Figure III.5.

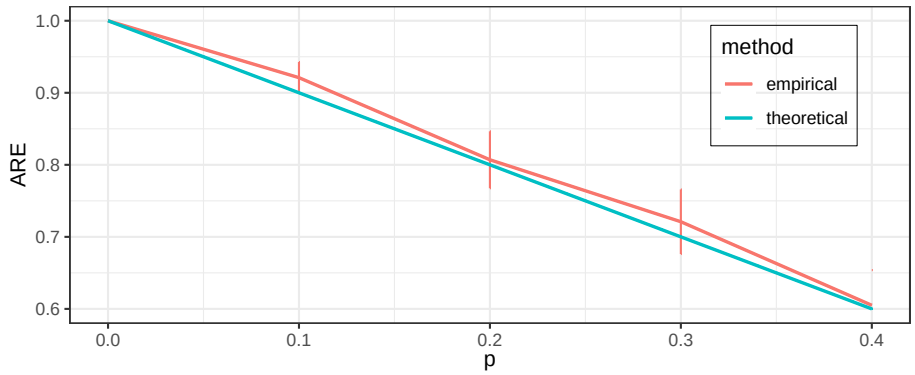

Figure III.5: Results of experiments with missing sensitive attributes. Empirical and theoretical curves of relative efficiency.

3.6 Summary and Discussion

In this work, we study the statistical impact of privacy and missingness on fairness under the task of learning fair machine learning models with noisy sensitive attributes. We define a restricted notion of asymptotic statistical efficiency in order to examine such impact. Quantitatively, the cost of noise induced by privacy or missingness on fairness generalizability is represented by a relative decline in statistical efficiency. The relative efficiency loss is interpretable: it explicitly depends on the privacy budget or missing proportion, subpopulation imbalance level, and a number of other problem-specific quantities. We validate and demonstrate the utility of our theory by a synthetic task of risk-parity linear regression with private group membership and a real data fairness-sensitive classification task with private or missing demographics.

Now we discuss a potential future research problem. For the sake of clarity, we consider $h \equiv 1$. Denote the loss vectors with regard to the true sensitive attribute

A and the noisy sensitive attribute Z, and the Markov transition matrix induced by the privacy mechanism Q (3.3.6) by

$$L_A(\theta) = \begin{bmatrix} \mathbb{E}\big[g(\theta; X, Y)|A = 1\big] \\ \mathbb{E}\big[g(\theta; X, Y)|A = 0\big] \end{bmatrix}, L_Z(\theta) = \begin{bmatrix} \mathbb{E}\big[g(\theta; X, Y)|Z = 1\big] \\ \mathbb{E}\big[g(\theta; X, Y)|Z = 0\big] \end{bmatrix}$$

and

$$M = \begin{bmatrix} 1-\gamma & \gamma \\ \gamma & 1-\gamma \end{bmatrix}.$$

Further, let $\mathbf{b} = (1, -1)^\top$. Noiseless, noisy, and debiased constraints are equivalent to each other at the population level in the way that $\mathbf{b}^\top L_A(\theta) = 0 \iff \mathbf{b}^\top L_Z(\theta) = 0 \iff \mathbf{b}^\top M^{-1} L_Z(\theta) = 0$. Consider their empirical counterparts, we note that $\mathbf{b}^\top \widehat{L}_{Z,n}(\theta) = 0 \iff \mathbf{b}^\top M^{-1} \widehat{L}_{Z,n}(\theta) = 0$. Combined with our theory, this empirical level equivalence of two constraints implies that using the inverse of the empirical transition matrix to match the noisy constraint to the noiseless constraint cannot improve the efficiency of the in-processing training procedure. Developing a principled in-processing method to increase the statistical efficiency is therefore an intriguing direction for future research.

CHAPTER IV

Distributionally Robust Performative Prediction

4.1 Introduction

In numerous fields where predictive analytics play an important role, decisions made on the basis of machine learning models do not just passively predict outcomes but actively influence future input data. Consider the domain of financial services, such as credit scoring and loan issuance, where a model's decision to grant or deny an application can affect the applicant's future financial behaviors and, consequently, the profile of future applicants. Similarly, in educational settings, the decision process for school admissions can shape the applicant pool, as those who are accepted often share their success strategies, indirectly influencing the preparation of future candidates. These examples highlight the study of performative prediction (Perdomo et al., 2020), a recent framework that facilitates a formal examination of learning in the presence of performative distribution shift resulting from deployed ML models.

Delving deeper into the formulations of performative prediction, the concept of a distribution map emerges as pivotal. This map characterizes the impact that a deployed ML model has on the underlying data distribution, which is a crucial element in navigating performative effects. The literature primarily revolves around the pursuit of performative stability (PS), a model which is optimal for the distribution it induces (Perdomo et al., 2020; Mendler-Dünner et al., 2020; Brown et al., 2022; Drusvyatskiy

and Xiao, 2023). However, a more ambitious target is the performative optimum (PO), which seeks the minimization of performative risk, the risk of the deployed model on the distribution it induces. Efficiently achieving PO typically necessitates a modeling of the distribution map (Miller et al., 2021; Izzo et al., 2021, 2022; Lin and Zrnic, 2023). Practically, the precise influence of a model on the data ecosystem is intricate and dynamic, making perfect specification an unattainable ideal.

In this chapter, we propose a distributionally robust performative prediction framework that aims to enhance robustness against a spectrum of distribution maps, thereby mitigating the issue of misspecification. Our contributions are summarized as follows:

- In Section 4.2, we formalize the DRPO concept, anchoring it within the performative prediction literature as a robust alternative;
- In Section 4.3, we provide theoretical insights into the efficacy of DRPO, demonstrating its resilience in the face of distribution map misspecification;
- In Section 4.4, we recast distributionally robust performative risk minimization as an augmented performative risk minimization problem, facilitating efficient optimization;
- In Section 4.5, we showcase DRPO's advatanges over conventional PO by empirical experiments.

The chapter concludes with a summary and discussion.

4.1.1 Related Work

Performative prediction. Performative prediction is an emerging framework for learning models that influence the data they intend to predict. The majority of research focuses on performative stability (Perdomo et al., 2020; Mendler-Dünner

et al., 2020; Brown et al., 2022; Ray et al., 2022; Li and Wai, 2022; Drusvyatskiy and Xiao, 2023; Dong et al., 2023), albeit with a few exceptions aiming at performative optimality. Miller et al. (2021) propose a two-stage plug-in method for finding the PO. Lin and Zrnic (2023) justify such plug-in method if the plug-in distribution map is wrong. Izzo et al. (2021) propose to find the PO by a parametric model of the distribution map. This method is extended by Izzo et al. (2022) to stateful performative prediction.

Distributionally robust optimization. DRO solves a stochastic optimization problem by optimizing under the worst-case scenario over an uncertainty set of probability distributions. Most popular DRO frameworks are based on φ-divergence (Hu and Hong, 2013; Bayraksan and Love, 2015; Gotoh et al., 2018; Gupta, 2019; Lam, 2019; Duchi et al., 2021) and Wasserstein distance (Gotoh et al., 2018; Mohajerin Esfahani and Kuhn, 2018; Chen and Paschalidis, 2018; Shafieezadeh-Abadeh et al., 2019; Blanchet and Murthy, 2019; Blanchet and Kang, 2021; Gao and Kleywegt, 2023). The existing DRO literature pays no attention to performative prediction except for Peet-Pare et al. (2022).

The works that are most pertinent to ours are Peet-Pare et al. (2022) and Lin and Zrnic (2023). Peet-Pare et al. (2022) study the repeated distributionally robust optimization algorithm, which repeatedly minimizes the distributionally robust risk at the induced distribution. They show such repeated training algorithm yields a distributionally robust performative stable (DRPS) solution, assuming conditions analogous to the validity of repeated risk minimization (Perdomo et al., 2020; Mendler-Dünner et al., 2020) in finding the performative stable (PS) solution. Nevertheless, the DRPO solution, which lies at the heart of the distributionally robust performative prediction problem, is not the subject of their study. Furthermore, they lacks theoretical guarantees regarding the proximity of the DRPS to either the PS or PO solution.

Lin and Zrnic (2023) argue that the PO with a misspecified nominal distribution

map can still reasonably approximate the true PO, as long as the misspecification level is not significant. This claim is also supported by our theory and experimental findings. Unlike them, we demonstrate that the DRPO is comparable to the PO if the misspecification is small, whereas the DRPO can offer substantial advantages over the PO if the misspecification is moderate to large. Moreover, the DRPO ensures reasonable performance for all distribution maps in an uncertainty collection surrounding the nominal distribution map, rather than just a single distribution map.

4.2 Methodology

4.2.1 Performative Prediction Essentials

Let Θ denote the (finite-dimensional) model parameter space, $\mathcal{Z}$ denote the data sample space, and $\mathcal{P}(\mathcal{Z})$ denote the set of probability measures supported on $\mathcal{Z}$. In performative prediction, we aim to find a $\theta \in \Theta$ that achieves low *performative risk*

$$\mathrm{PR}_{\mathrm{true}}(\theta) = \mathbb{E}_{Z\sim\mathcal{D}_{\mathrm{true}}(\theta)}[\ell(Z;\theta)], \tag{4.2.1}$$

where $\ell : \mathcal{Z} \times \Theta \to \mathbb{R}$ is the (known) loss function, and $\mathcal{D}_{\mathrm{true}} : \Theta \to \mathcal{P}(\mathcal{Z})$ is the *true distribution map*. The *true performative optimum* (true PO) $\theta_{\mathrm{PO,true}}$ is known to minimize the (true) *performative risk*:

$$\theta_{\mathrm{PO,true}} \in \arg\min_{\theta\in\Theta} \left\{\mathrm{PR}_{\mathrm{true}}(\theta) = \mathbb{E}_{Z\sim\mathcal{D}_{\mathrm{true}}(\theta)}[\ell(Z;\theta)]\right\}. \tag{4.2.2}$$

It is easy to see that if we *know* the true distribution map, then we can evaluate the true performative risk and find the true performative optimum well up to some finite sample error which is negligible as the sample size goes to infinity. However, the true map $\mathcal{D}_{\mathrm{true}}(\cdot)$ is *unknown* in general, thus posing a significant obstacle in the pursuit of evaluating and optimizing the true performative risk.

To enable the optimization of performative risk, it is necessary to have a known *nominal distribution map* $\mathcal{D}(\cdot)$ that is believed to closely approximate the unknown true distribution map $\mathcal{D}_{\text{true}}(\cdot)$. Then one can find the *performative optimum* (PO) by minimizing the *nominal performative risk*:

$$\theta_{\text{PO}} \in \arg\min_{\theta\in\Theta} \left\{ \text{PR}(\theta) = \mathbb{E}_{Z\sim\mathcal{D}(\theta)}[\ell(Z;\theta)] \right\}. \tag{4.2.3}$$

Because the distribution map is inevitably *misspecified*, $\mathcal{D}(\cdot) \neq \mathcal{D}_{\text{true}}(\cdot)$, the true performative risk is generally not minimized by θ_{PO}. Therefore, we treat (4.2.3) as a solution concept to approximately solve the true performative risk minimization problem (4.2.2), and refer (4.2.3) to *standard performative prediction*.

We provide several illustrative instances of potential sources that may lead to the misspecification of distribution map, *i.e.*, $\mathcal{D}(\cdot) \neq \mathcal{D}_{\text{true}}(\cdot)$.

Modeling error. The modeling of $\mathcal{D}(\cdot)$ can be either a *deterministic model* of explicit form or a *statistical model* with model parameters to be estimated. The misspecification of $\mathcal{D}(\cdot)$ may stem from modeling error.

Example IV.1 (Strategic classification)**.** *Let $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$ where $\mathcal{X}$ is the feature space and $\mathcal{Y}$ is the label space, so that we are in the supervised learning regime. Strategic classification relies on a working model of individual's data manipulation strategy:*

$$\Delta_\theta(x) = \arg\max_{x'}\{u(x') - c(x,x')\}, \tag{4.2.4}$$

where $\Delta_\theta(\cdot)$, $u(\cdot)$, and $c(\cdot,\cdot)$ are known as individual's best response function, utility function, and cost function, respectively. The best response function can be lifted to the measurable space of $\mathcal{X}\times\mathcal{Y}$ so that we have the response map

$$T_\theta\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} \Delta_\theta(x) \\ y \end{bmatrix} = \begin{bmatrix} \arg\max_{x'}\{u(x') - c(x,x')\} \\ y \end{bmatrix}.$$

The nominal distribution map $\mathcal{D}(\cdot)$ is fully characterized by $T_\theta(\cdot)$ and the sampling distribution of $\mathcal{D}_{\text{true}}(\mathbf{0})$ (see details in Section 4.4.1). In this example, the distribution map $\mathcal{D}(\cdot)$ could be misspecified because the individual's utility function and cost function could be misspecified.

Example IV.2 (Location family)**.** *Location family postulates a translation model:*

$$Z \sim \mathcal{D}(\theta) \iff Z \stackrel{d}{=} Z_0 + A\theta \text{ where } Z_0 \sim \mathcal{D}_{\text{true}}(\mathbf{0}),$$

where $A \in \mathbb{R}^{\dim(\mathcal{Z})\times\dim(\Theta)}$ is unknown and therefore must be estimated. If we observe the sampling distributions of $\mathcal{D}(\theta_0), \mathcal{D}(\theta_1) \ldots, \mathcal{D}(\theta_K)$, then A is partially identified up to a linear subspace of $\mathbb{R}^{\dim(\mathcal{Z})\times\dim(\Theta)}$:

$$A \in \{M \mid M(\theta_k - \theta_0) = \mu_k - \mu_0 \text{ for } k \in [K]\},$$

where μ_k is the mean of $\mathcal{D}(\theta_k)$. In this example, the distribution map $\mathcal{D}(\cdot)$ can be misspecified because the model parameter A is only partially identifiable.

Distribution shift. The misspecification of $\mathcal{D}(\cdot)$ can be even worse in the presence of distribution shift. Consider the training and test environments have different distribution maps, $\mathcal{D}_{\text{train}}(\cdot)$ and $\mathcal{D}_{\text{test}}(\cdot)$, respectively. Even if we correctly specify the nominal distribution $\mathcal{D}(\cdot) = \mathcal{D}_{\text{train}}(\cdot)$, $\mathcal{D}(\cdot)$ can still be misspecified for $\mathcal{D}_{\text{true}}(\cdot) = \mathcal{D}_{\text{test}}(\cdot)$ due to $\mathcal{D}_{\text{train}}(\cdot) \neq \mathcal{D}_{\text{test}}(\cdot)$.

Example IV.3 (Disparate impacts and fairness)**.** *A population is comprised of majority and minority subpopulations (e.g., by race or gender). The population distribution map is a mixture of the subpopulation distribution maps: $\mathcal{D}_{\text{pop}}(\theta) \stackrel{d}{=} \gamma\mathcal{D}_{\text{maj}}(\theta) + (1-\gamma)\mathcal{D}_{\text{min}}(\theta)$. In fair machine learning, a theme is to check whether an ML model has disparate impacts on different subpopulations or biased against the minority. Suppose that we work with $\mathcal{D}(\cdot) = \mathcal{D}_{\text{pop}}(\cdot)$ and target at the minority $\mathcal{D}_{\text{true}}(\cdot) = \mathcal{D}_{\text{min}}(\cdot)$.*

However, the population distribution map and the minority distribution map may differ. In this example, the distribution map $\mathcal{D}(\cdot)$ is misspecified because of subpopulation shift.

4.2.2 Distributionally Robust Performative Prediction

From an intuitive perspective, the PO solution (4.2.3) has the potential to achieve low performative risk $\mathrm{PR}_{\mathrm{true}}(\theta_{\mathrm{PO}})$ when the nominal distribution map $\mathcal{D}(\cdot)$ closely aligns with the true distribution map $\mathcal{D}_{\mathrm{true}}(\cdot)$. However, $\mathcal{D}(\cdot)$ and $\mathcal{D}_{\mathrm{true}}(\cdot)$ may be quite different. In such cases, the PO solution may incur high performative risk. To address this issue, we propose a distributionally robust formulation for performative prediction, where we explicitly incorporate into the learning phase the consideration that the true distribution map $\mathcal{D}_{\mathrm{true}}(\cdot)$ is different from the nominal one $\mathcal{D}(\cdot)$.

Let $D(\cdot\|\cdot)$ be the KL divergence, *i.e.*,

$$D(Q\|P) = \int \varphi\left(\frac{dQ}{dP}\right) dP = \int \log\left(\frac{dQ}{dP}\right) dQ,$$

where $\varphi(t) = t\log t$ for any $t > 0$. Here dQ/dP is the Radon–Nikodym derivative, and we implicitly require the probability measure Q to be absolutely continuous with respect to P. With the KL divergence, we can define a family of distribution maps around the nominal distribution map. Specifically, the *uncertainty collection* around $\mathcal{D}$ with *critical radius* ρ is defined as

$$\mathcal{U}(\mathcal{D}) = \{\widetilde{\mathcal{D}} : \Theta \to \mathcal{P}(\mathcal{Z}) \mid D(\widetilde{\mathcal{D}}(\theta)\|\mathcal{D}(\theta)) \le \rho, \forall\theta \in \Theta\}.$$

The radius ρ reflects the the magnitude of shifts in distribution map we seek to be robust to. We remark that the value of ρ can be prescribed or be selected in data-driven ways. We postpone the discussion on critical radius calibration to Section 4.4.4.

Definition 4.2.1 (Distributionally robust performative risk)**.** The distributionally robust performative risk with the uncertainty collection $\mathcal{U}(\mathcal{D})$ is defined as

$$\mathrm{DRPR}(\theta) = \sup_{\widetilde{\mathcal{D}}:\widetilde{\mathcal{D}}\in\mathcal{U}(\mathcal{D})} \mathbb{E}_{Z\sim\widetilde{\mathcal{D}}(\theta)}[\ell(Z;\theta)]. \tag{4.2.5}$$

In other words, the distributionally robust performative risk $\mathrm{DRPR}(\theta)$ measures the worst possible performative risk incurred by the model parameterized by θ among the collection of all alternative distribution maps that are ρ-close to the nominal distribution map $\mathcal{D}$. With this intuition, it is natural to define an alternative solution concept which minimizes (4.2.5).

Definition 4.2.2 (Distributionally robust performative optimum)**.** The distributionally robust performative optimum (DRPO) is defined as

$$\theta_{\mathrm{DRPO}} \in \arg\min_{\theta\in\Theta} \mathrm{DRPR}(\theta). \tag{4.2.6}$$

We refer the method (4.2.6) to *distributionally robust performative prediction*. When comparing (4.2.6) and (4.2.3), we view the DRPO (4.2.6) as a competing solution concept to the PO (4.2.3), because both of them aim for achieving low performance risk (4.2.1).

4.2.3 Generalization Principle of DRPO

Distributionally robust performative prediction asks to not only perform well on a fixed performative prediction problem (parameterized by the distribution map $\mathcal{D}$), but simultaneously for a range of performative prediction problems, each determined by a distribution map in an uncertainty collection $\mathcal{U}$. This results in more robust solutions, that is, those DRPOs which are robust to misspecification of distribution map. The uncertainty collection plays a key role: it implicitly defines the induced notion of robustness.

Distributionally robust performative prediction yields a natural approach for certifying out-of-sample performance, which is summarized by the following principle.

Proposition 4.2.1 (Generalization principle of distributionally robust performative prediction)**.** *Suppose that the uncertainty collection $\mathcal{U}$ contains the true distribution map $\mathcal{D}_{\text{true}}$, then the true performative risk is bounded by the distribution robust performative risk:*

$$\text{PR}_{\text{true}}(\theta) \leq \text{DRPR}(\theta) \tag{4.2.7}$$

for any $\theta \in \Theta$. In consequence, we have

$$\text{PR}_{\text{true}}(\theta_{\text{DRPO}}) \leq \text{DRPR}(\theta_{\text{DRPO}}).$$

Essentially, if the uncertainty collection $\mathcal{U}$ is chosen appropriately, the corresponding DR performative risk gives an upper bound for the true performative risk, and thus the DRPO enjoys provable guarantees on its incurred performative risk. Furthermore, the generalization principle can be extended to non-deterministic statement: if $\mathcal{U}$ contains $\mathcal{D}$ with high probability $1-\delta$, then the generalization bound (4.2.7) holds with high probability $1-\delta$. This suggests that we may choose uncertainty collection $\mathcal{U}$ in some data-driven ways.

4.2.4 Benefits of DRPO: A Toy Example

A toy example is employed to facilitate a conceptual understanding of the advantages of DRPO over PO in relation to achieving improved control over the worst-case performative risk.

Let $\mathcal{Z} = \mathbb{R}$, $\Theta = [-1, 1]$, and $\ell(z; \theta) = \theta z$. Let the nominal distribution map be $\mathcal{D}(\theta) = \mathcal{N}(f(\theta), \sigma^2)$ for some $f : [-1, 1] \to \mathbb{R}$ and $\sigma^2 > 0$. Then the nominal

performative risk is

$$\mathrm{PR}(\theta) = \mathbb{E}_{Z\sim\mathcal{N}(f(\theta),\sigma^2)}[\ell(Z;\theta)] = \theta f(\theta).$$

Regularization effect. With the dual formula given in Section 4.3.1, one can derive the distributionally robust performative risk directly:

$$\mathrm{DRPR}(\theta) = \mathrm{PR}(\theta) + \sqrt{\rho}\,\mathrm{Penalty}(\theta),$$

where the penalty function $\mathrm{Penalty}(\theta) = \sqrt{2\sigma^2}\,|\theta|$ penalizes the deviation of θ from the origin 0, and the critical radius ρ tunes the level of regularization. That is to say, in this toy example, the distributionally robust performative risk minimization problem is essentially a L_1-*regularized* performative risk minimization problem.

Better worst-case control. To be more concrete, we let $f(\theta) = a_1\theta + a_0$ for some $a_1, a_0 > 0$. For any $\widetilde{\mathcal{D}} \in \mathcal{U}(\mathcal{D})$, let the performative risk of $\widetilde{\mathcal{D}}$ be

$$\mathrm{PR}_{\widetilde{\mathcal{D}}}(\theta) = \mathbb{E}_{Z\sim\widetilde{\mathcal{D}}(\theta)}[\ell(Z;\theta)].$$

If $\widetilde{\mathcal{D}}$ is the true distribution map, then $\mathrm{PR}_{\widetilde{\mathcal{D}}}(\theta)$ is the true performative risk that we incur. Through direct calculation (see details in Appendix C.1), one can show that θ_{DRPO} is more robust than θ_{PO} in the sense of *worst-case performative risk* control, that is,

$$\sup_{\widetilde{\mathcal{D}}\in\mathcal{U}(\mathcal{D})} \mathrm{PR}_{\widetilde{\mathcal{D}}}(\theta_{\mathrm{PO}}) \geq \sup_{\widetilde{\mathcal{D}}\in\mathcal{U}(\mathcal{D})} \mathrm{PR}_{\widetilde{\mathcal{D}}}(\theta_{\mathrm{DRPO}}) + \frac{\rho\sigma^2}{2a_1}$$

for any fixed $\rho \leq \frac{a_0^2}{2\sigma^2}$. This example shows that the worst-case performative risk incurred by the PO can be *arbitrarily larger than* that incurred by the DRPO, as $a_1 \to 0$. The message that DRPO offers certain advantages over PO in terms of mitigating worst-case performative risk is also supported by the empirical results

presented in Section 4.5.

4.3 Theory

4.3.1 Strong Duality of DRPR

The evaluation of distributionally robust performative risk $\mathrm{DRPR}(\theta)$ given in (4.2.5) involves an infinite dimensional maximization problem which is generally intractable. Fortunately, it is in fact equivalent to a minimization problem over a single dual variable. This is given by the following strong duality result.

Proposition 4.3.1 (Strong duality of DRPR)**.** *For any $\theta \in \Theta$, we have*

$$\mathrm{DRPR}(\theta) = \inf_{\mu \geq 0} \left\{ \mu \log \mathbb{E}_{Z \sim \mathcal{D}(\theta)} \left[e^{\ell(Z;\theta)/\mu} \right] + \mu\rho \right\}. \tag{4.3.1}$$

The dual reformulation (4.3.1) will be served as the cornerstone of developing algorithms for finding the DRPO in Section 4.4.

4.3.2 Worst-case Distribution Map

As a byproduct of the proof of Proposition 4.3.1, here we provide a characterization of the worst-case distribution map which attains the supremum in (4.2.5).

Proposition 4.3.2 (Worst-case distribution map)**.** *Suppose the problem* (4.3.1) *is solved by the unique $\mu^\star(\theta) > 0$ for any $\theta \in \Theta$. We define a distribution map $\widetilde{\mathcal{D}}$ satisfying the density ratio equation*

$$\frac{d\widetilde{\mathcal{D}}(\theta)}{d\mathcal{D}(\theta)} = \frac{e^{\ell(Z;\theta)/\mu^\star(\theta)}}{\mathbb{E}_{Z \sim \mathcal{D}(\theta)}\left[e^{\ell(Z;\theta)/\mu^\star(\theta)}\right]}, \quad \forall \theta \in \Theta. \tag{4.3.2}$$

Then we have that $\widetilde{\mathcal{D}}$ is the unique worst-case distribution map, that is,

$$\widetilde{\mathcal{D}}(\theta) = \underset{\widetilde{\mathcal{D}}(\theta):\widetilde{\mathcal{D}}\in\mathcal{U}(\mathcal{D})}{\arg\sup}\ \mathbb{E}_{Z\sim\widetilde{\mathcal{D}}(\theta)}[\ell(Z;\theta)]. \tag{4.3.3}$$

Proposition 4.3.2 shows that the worst-case distribution map $\widetilde{\mathcal{D}}$ is an *exponentially tilted* distribution map with respect to the nominal distribution map $\mathcal{D}$, where $\widetilde{\mathcal{D}}$ puts more weights on the high ends. This implies that the DR performative prediction regulates the right tail of the performative losses. The regulation effects are further demonstrated by Figure IV.2 in Experiment 4.5.1.

4.3.3 Excess Risk Bound

Now we are interested in bounding the excess risk:

$$\begin{aligned}\mathcal{E}(\hat{\theta}) &= \mathrm{PR}_{\mathrm{true}}(\hat{\theta}) - \min_{\theta\in\Theta}\mathrm{PR}_{\mathrm{true}}(\theta)\\ &= \mathrm{PR}_{\mathrm{true}}(\hat{\theta}) - \mathrm{PR}_{\mathrm{true}}(\theta_{\mathrm{PO,true}})\end{aligned} \tag{4.3.4}$$

where $\hat{\theta}$ is an approximate solution to the true PO $\theta_{\mathrm{PO,true}}$, which could particually be θ_{PO} and θ_{DRPO}. The excess risk (4.3.4) captures the true performance of $\hat{\theta}$ relative to the oracle performative optimum $\theta_{\mathrm{PO,true}}$, providing a direct measurement of the suboptimality of $\hat{\theta}$ in terms of performative risk. Now we show the excess risk bounds of the PO solution and the DRPO, $\mathcal{E}(\theta_{\mathrm{PO}})$ and $\mathcal{E}(\theta_{\mathrm{DRPO}})$, and compare them.

Proposition 4.3.3 (Excess risk bound of the PO)**.** *Suppose that we have bounded loss function $|\ell(z;\theta)| \le B$ for any $z \in \mathcal{Z}, \theta \in \Theta$ and some $B > 0$. Then we have*

$$\mathcal{E}(\theta_{\mathrm{PO}}) \le \sqrt{2}B \sup_{\theta\in\Theta}\sqrt{D(\mathcal{D}_{\mathrm{true}}(\theta)\|\mathcal{D}(\theta))}. \tag{4.3.5}$$

Proposition 4.3.4 (Excess risk bound of the DRPO)**.** *Suppose that*

$$D(\mathcal{D}_{\text{true}}(\theta_{\text{DRPO}})\|\mathcal{D}(\theta_{\text{DRPO}})) \leq \rho.$$

Then we have

$$\mathcal{E}(\theta_{\text{DRPO}}) \leq \sqrt{\rho \operatorname{Var}_{Z\sim\mathcal{D}(\theta_{\text{PO,true}})}[\ell(Z;\theta_{\text{PO,true}})]} + o(\sqrt{\rho}). \tag{4.3.6}$$

Comparing Proposition 4.3.3 and 4.3.4, we see that the excess risk bound of the DRPO can be localized to the true PO while the excess risk bound of the PO is entangled with the full parametric space Θ. Although we are comparing two upper bounds which can be not tight enough, the comparison sheds lights to the potential benefits of using DRPO over PO solution. Even if in the case of no significant improvement of using DRPO over PO solution, the excess risks of them are comparable, thus doing no harm. Our insight has been verified through a toy example in Section 4.2.4 and as well experimental results in Section 4.5.

4.3.4 Generalized Excess Risk Bound

We upper bound $\mathcal{E}(\theta_{\text{DRPO}})$ when the uncertainty collection doesn't cover the true distribution map. To be specific, we analyze the excess risk bound of the DRPO when $D(\mathcal{D}_{\text{true}}(\theta_{\text{DRPO}})\|\mathcal{D}(\theta_{\text{DRPO}})) > \rho$.

Proposition 4.3.5 (Excess risk bound of the DRPO in general)**.** *Suppose that we have bounded loss function $|\ell(z;\theta)| \leq B$ for any $z \in \mathcal{Z}, \theta \in \Theta$ and some $B > 0$. Then we have*

$$\begin{aligned}\mathcal{E}(\theta_{\text{DRPO}}) \leq & \sqrt{\rho \operatorname{Var}_{Z\sim\mathcal{D}(\theta_{\text{PO,true}})}[\ell(Z;\theta_{\text{PO,true}})]} + \\ & \sqrt{2}B \inf_{P:D(P\|\mathcal{D}(\theta_{\text{DRPO}}))\leq\rho} \sqrt{D(\mathcal{D}_{\text{true}}(\theta_{\text{DRPO}})\|P)} + o(\sqrt{\rho}).\end{aligned} \tag{4.3.7}$$

Comparing (4.3.7) to (4.3.6), we have an additional term in the excess risk bound which accommodates and accounts for the misspecification of uncertainty set around $\mathcal{D}(\theta_{\mathrm{DRPO}})$, which doesn't necessarily cover $\mathcal{D}_{\mathrm{true}}(\theta_{\mathrm{DRPO}})$. Furthermore, it is not difficult to see that if $D(\mathcal{D}_{\mathrm{true}}(\theta_{\mathrm{DRPO}})\|\mathcal{D}(\theta_{\mathrm{DRPO}})) \leq \rho$, then the last infimum term in the upper bound of (4.3.7) vanishes and (4.3.7) reduces to (4.3.6). Therefore, Proposition 4.3.5 provides a generalized excess risk bound for the DRPO than Proposition 4.3.4.

4.4 Algorithms

4.4.1 Performative Risk Minimization

We recall a standard algorithm for performative risk minimization when the nominal distribution map $\mathcal{D}(\cdot)$ is modeled by a *response map* $T_\theta : \mathcal{Z} \to \mathcal{Z}$ and samples from the *base distribution* $\mathcal{D}_{\mathrm{true}}(\mathbf{0})$:

$$Z \sim \mathcal{D}(\theta) \iff Z \overset{d}{=} T_\theta(Z_0) \text{ where } Z_0 \sim \mathcal{D}_{\mathrm{true}}(\mathbf{0}).$$

This is a popular model for distribution map including strategic classification (see Example IV.1) and location family (see Example IV.2) as prominent examples.

It is not hard to see that the nominal distribution map $\mathcal{D}(\cdot)$ is fully characterized by the response map T_θ and the base measure $\mathcal{D}_{\mathrm{true}}(\mathbf{0})$ because

$$\mathcal{D}(\theta) \overset{d}{=} T_{\theta\sharp}(\mathcal{D}_{\mathrm{true}}(\mathbf{0})) \text{ for any } \theta \in \Theta,$$

that is, the measure $\mathcal{D}(\theta)$ is the *pushforward measure* of $\mathcal{D}_{\mathrm{true}}(\mathbf{0})$ under the *transport*

map T_θ. The performative risk can then be reformulated as

$$\begin{aligned}\mathrm{PR}(\theta) &= \mathbb{E}_{Z\sim\mathcal{D}(\theta)}[\ell(Z;\theta)] \\ &= \mathbb{E}_{Z\sim T_{\theta\sharp}(\mathcal{D}_{\mathrm{true}}(\mathbf{0}))}[\ell(Z;\theta)] \\ &= \mathbb{E}_{Z\sim\mathcal{D}_{\mathrm{true}}(\mathbf{0})}[\ell(T_\theta(Z);\theta)].\end{aligned}$$

With the last formula, the performative risk minimization problem becomes a standard stochastic optimization problem, which can be solved efficiently by any popular stochastic optimization algorithm, *e.g.*, sample average approximation (SAA, Kleywegt et al. (2002)).

We refer to Section 4.1.1 for a summary of existing performative risk minimization algorithms. In the following subsections, we will see that any off-the-shelf algorithms for finding the PO can be utilized as an *intermediate algorithm* for finding the DRPO.

4.4.2 Distributionally Robust Performative Risk Minimization

By the strong duality in Proposition 4.3.1, DR performative risk minimization is equivalent to the following optimization problem jointly over (θ, μ):

$$\begin{gathered}\min_\theta \mathrm{DRPR}(\theta) \iff \min_{\theta\in\Theta}\inf_{\mu\geq 0}\psi(\theta,\mu) \\ \text{where } \psi(\theta,\mu) = \mu\log\mathbb{E}_{Z\sim\mathcal{D}(\theta)}\left[e^{\ell(Z;\theta)/\mu}\right] + \mu\rho.\end{gathered} \tag{4.4.1}$$

This suggests an alternating minimization algorithm, summarized in Algorithm 2, where we learn θ_{DRPO} by fixing μ and minimizing on θ and then fixing θ and minimizing on μ alternatively until convergence.

The step of minimizing on θ with fixed μ (Line 4 in Algorithm 2) is itself a performative risk minimization problem, which can be solved by any suitable performative risk minimization algorithm. The step of minimizing on μ with fixed θ (Line 5 in Algorithm 2) can be solved by the line search or the Newton–Raphson method since

Algorithm 2 DR Performative Risk Minimization

Input: radius ρ, nominal distribution map $\mathcal{D}(\theta)$
Initialize μ
while μ has not converged **do**
    Update $\theta \leftarrow \arg\min_{\theta\in\Theta} \left\{\mathbb{E}_{Z\sim\mathcal{D}(\theta)}\left[e^{\ell(Z;\theta)/\mu}\right]\right\}$
    Update $\mu \leftarrow \arg\min_{\mu\geq 0} \{\psi(\theta,\mu)\}$ (ψ in (4.4.1))
end while
Return: θ

Algorithm 3 Tilted Performative Risk Minimization

Input: tilt α, nominal distribution map $\mathcal{D}(\theta)$
Update $\theta \leftarrow \arg\min_{\theta\in\Theta} \left\{\mathbb{E}_{Z\sim\mathcal{D}(\theta)}\left[e^{\alpha\ell(Z;\theta)}\right]\right\}$
Return: θ

$\psi(\theta,\mu)$ is convex in μ. Overall, the total cost of Algorithm 2 is therefore comparable to that of the performative risk minimization algorithm used in the intermediate step.

4.4.3 Tilted Performative Risk Minimization

Treating ρ as a hyperparameter which can be tuned by a practitioner, the solution of the dual problem (4.4.1) can be denoted by $(\theta^\star(\rho), \mu^\star(\rho))$. One can show that $\mu^\star(\rho)$ is a decreasing function of ρ. An intuition is that as $\mu \to \infty$, we have

$$\begin{aligned}
&\arg\min_{\theta\in\Theta} \left\{\mu \log \mathbb{E}_{Z\sim\mathcal{D}(\theta)}\left[e^{\ell(Z;\theta)/\mu}\right] + \mu\rho\right\} \\
\approx\ &\arg\min_{\theta\in\Theta} \left\{\mu \log \mathbb{E}_{Z\sim\mathcal{D}(\theta)}\left[\frac{\ell(Z;\theta)}{\mu} + 1\right] + \mu\rho\right\} \\
\approx\ &\arg\min_{\theta\in\Theta} \left\{\mathbb{E}_{Z\sim\mathcal{D}(\theta)}[\ell(Z;\theta)]\right\},
\end{aligned}$$

which reduces to the original performative risk minimization problem, or the distributionally robust performative risk minimization problem with $\rho = 0$.

With this motivation, instead of tuning ρ, we can tune μ (or the inverse of it, denoted by $\alpha = \mu^{-1}$) and solve the *α-tilted performative risk minimization* problem:

$$\theta_{\mathrm{TPO}} \in \arg\min_{\theta\in\Theta} \left\{\mathrm{TPR}(\theta) = \mathbb{E}_{Z\sim\mathcal{D}(\theta)}\left[e^{\alpha\ell(Z;\theta)}\right]\right\}, \tag{4.4.2}$$

where $\mathrm{TPR}(\theta)$ stands for the tilted performative risk and θ_{TPO} is the *tilted performative optimum*, that is, the performative optimum of the tilted problem. In order to have stronger distributional robustness property, we tune α to be larger. Given the correspondence $\mu^\star(\rho)$, we should have θ_{TPO} with α equals θ_{DRPO} with $\rho = (\mu^\star)^{-1}(1/\alpha)$. Therefore, the tilted performative risk minimization *implicitly* solves a corresponding DR performative risk minimization problem.

Finally, we remark that exponential tilting is a statistical method that has been around at least since the exponential family (Keener, 2010) was first invented. More recently, it has been applied to operation research (Ahmadi-Javid, 2012) and machine learning (Li et al., 2020, 2023).

4.4.4 Calibration of Critical Radius

The performance of the DRPO is contingent on the uncertainty collection radius ρ, which is typically difficult to choose a priori without additional information. The greater the value of ρ, the higher the level of distributional robustness, and thus the greater the tolerance for distribution map misspecification. Therefore, the selection of ρ reflects a practitioner's risk-aversion preference. In this subsection, we present two simple, yet effective calibration techniques for selecting ρ.

Post-fitting calibration. Without additional information, we can only hope to be robust to a prescribed set of distribution maps, say Ξ. The assumption of Proposition 4.3.4 requires only the uncertainty collection at the DRPO, which reduces to an uncertainty ball centered at $\mathcal{D}(\theta_{\mathrm{DRPO}})$, encompassing the true distribution $\mathcal{D}_{\mathrm{true}}(\theta_{\mathrm{DRPO}})$. Therefore, in order to provide a provable guarantee for all distribution maps in Ξ, the radius ρ can be chosen based on the following criterion:

$$\rho_{\mathrm{cal}} = \arg\min_{\rho \geq 0} \left\{ \max_{\mathcal{D}_{\mathrm{true}} \in \Xi} \widehat{D}(\mathcal{D}_{\mathrm{true}} \| \mathcal{D}) \leq \rho \right\},$$

where

$$\widehat{D}(\mathcal{D}_{\text{true}}\|\mathcal{D}) := \widehat{D}(\mathcal{D}_{\text{true}}(\theta_{\text{DRPO}}(\rho))\|\mathcal{D}(\theta_{\text{DRPO}}(\rho))).$$

Here $\widehat{D}$ is the estimated KL divergence, and $\theta_{\text{DRPO}}(\rho)$ is indexed by the radius ρ to highlight its dependence on ρ as a tuning parameter to be calibrated. We implement the post-fitting calibration approach in Section 4.5.1.

Calibration set. With additional information, such as a small set of calibration data, we can pick ρ (or α if we use Algorithm 3) by evaluating the performance of $\theta_{\text{DRPO}}(\rho)$ on the calibration set. Consider Example IV.3 where the training and test distribution maps may differ, we can conduct the following grid searching procedure to choose ρ: 1) for a candidate set $\mathcal{C}$ of ρ's, we compute $\{\theta_{\text{DRPO}}(\rho) : \rho \in \mathcal{C}\}$ under $\mathcal{D}_{\text{train}}$; 2) we obtain a few calibrating samples from $\mathcal{D}_{\text{test}}$, on which we evaluate the performance of $\theta_{\text{DRPO}}(\rho)$; 3) we select $\rho_{\text{cal}} \in \mathcal{C}$ with the best calibration set performance. The calibration set approach is implemented in Section 4.5.3.

To conclude this subsection, we discuss the computational costs of the proposed calibration methods. For general problems, these calibration methods necessitate a grid search, which may be computationally expensive. Fortunately, for specific problems (for example, experiments in Section 4.5.1), we can exploit the decreasing nature of $\widehat{D}(\mathcal{D}_{\text{true}}(\theta_{\text{DRPO}}(\rho))\|\mathcal{D}(\theta_{\text{DRPO}}(\rho)))$ as a function in ρ. As a result, we can use bisection search rather than grid search to significantly reduce computational costs.

4.5 Experiments

We revisit Examples IV.1, IV.2, and IV.3 and compare the PO and the DRPO empirically. For a particular (true) distribution map, either the PO or the DRPO may have the potential to outperform the other in terms of the performative risk evaluated at this particular distribution map. In contrast to the PO, however, the DRPO aims to guarantee reasonable performance for *all* distribution maps in an uncertainty

collection around the nominal distribution map. To ensure the performance of the DRPO on a (set of) specific distribution map(s), the radius ρ (or the tilt α) must be calibrated. Lastly, each shaded region in figures below shows the curve's standard error of the mean from 30 trials.

4.5.1 Strategic Classification with Misspecified Cost Function

In reference to Example IV.1, we examine strategic classification involving a cost function that is misspecified. The experimental setup resembles that in Perdomo et al. (2020). The task is credit scoring, specifically predicting debt default. Individuals strategically manipulate their features to obtain a favorable classification.

Consider an instantiation of the response map (4.2.4) such that $u(x) = -\theta^\top x$ and $c(x, x') = \frac{1}{2\epsilon} \times \|x_{\text{strat}} - x'_{\text{strat}}\|_2^2 + \infty \times \|x_{\text{non-strat}} - x'_{\text{non-strat}}\|_2^2$. Without loss of generality, let the first m features be strategic features and the last $d - m$ features be non-strategic features. Let $B = \text{diag}(1, \ldots, 1, 0, \ldots, 0) \in \mathbb{R}^{d \times d}$ where the first m diagonal elements are 1's and the others are 0's. Then the best response function is $\Delta_\theta(x) = x - \epsilon B\theta$.

For the base distribution $\mathcal{D}(\mathbf{0})$, we use a class-balanced subset of a Kaggle credit scoring dataset (Kaggle, 2012). Features encompass an individual's historical data, including their monthly income and credit line count. Labels are binary where the value of 1 represents a default on a debt and 0 otherwise. There are a total of 3 strategic features and 6 non-strategic features. We use logistic model for the classifier and the cross-entropy loss with L_2-regularization for the loss function ℓ.

Consider the cost function is misspecified by its performativity level ϵ. We specify the cost function with the nominal performativity level $\epsilon = 0.5$. However, the true performativity level ϵ_{true} might not be 0.5, but in the range of $[0.5 - 0.5\eta, 0.5 + 0.5\eta]$ for some $\eta \geq 0$.

The top left plot of Figure IV.1 shows performative risk incurred by the PO

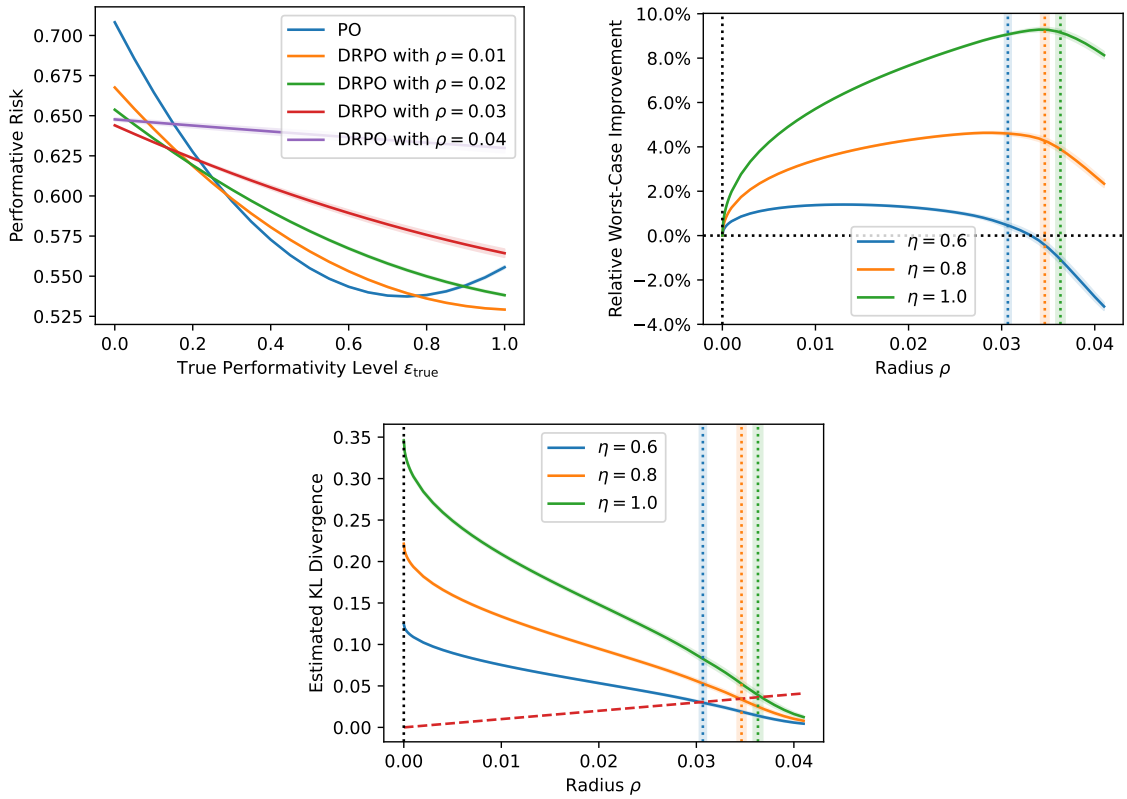

Figure IV.1: Results of Experiment 4.5.1. Top left: performative risk incurred by the PO and the DRPO's with various radius ρ's. Top right: relative improvement in worst-case performative risk of the DRPO to the PO as the radius ρ increases, for different range of misspecification η's. Bottom: radius ρ versus estimated KL divergence between $\mathcal{D}_{\text{true}}(\theta_{\text{DRPO}}(\rho))$ and $\mathcal{D}(\theta_{\text{DRPO}}(\rho))$, where vertical bands indicate the calibrated radius ρ_{cal}'s.

and the DRPO's with various radius ρ's. Note that the PO can be understood as the DRPO with $\rho = 0$. As ρ increases, the DRPO aims to achieve more uniform performance across a wider range of ϵ_{true}. The top right plot of Figure IV.1 shows relative improvement in worst-case performative risk[1] of the DRPO to the PO as the radius ρ increases, for different range of misspecification η's. Curves positioned above the horizontal dotted line indicate that the DRPO outperform the PO in terms of worst-case performative risk. When there is a larger range of misspecification η, the DRPO has greater potential to beat the PO. The bottom plot of Figure IV.1

[1]The relative worst-case improvement of θ to θ_{PO} is defined by $\frac{\max_{\epsilon_{\text{true}}} \text{PR}_{\text{true}}(\theta_{\text{PO}}) - \max_{\epsilon_{\text{true}}} \text{PR}_{\text{true}}(\theta)}{\max_{\epsilon_{\text{true}}} \text{PR}_{\text{true}}(\theta_{\text{PO}})} \times 100\%$.

demonstrates the post-fitting calibration of radius ρ_{cal} described in Section 4.4.4. The vertical bands indicate the calibrated radius ρ_{cal}'s, which lead to the DRPO's with sound relative worst-case improvement and avoid overconservative solutions, as depicted by the corresponding bands in the middle plot of Figure IV.1.

Figure IV.2 shows histogram of performative loss for the PO, the DRPO with $\rho = 0.02$, and the DRPO with $\rho = 0.04$, under the setup of Experiment 4.5.1 with $\epsilon_{\text{true}} = 0.5$. These plots displayed in a left-to-right-to-bottom manner demonstrate that the DRPO regulates the right tail of the performative losses. Moreover, as the radius ρ increases, there is a corresponding increase in the degree of regulation effects.

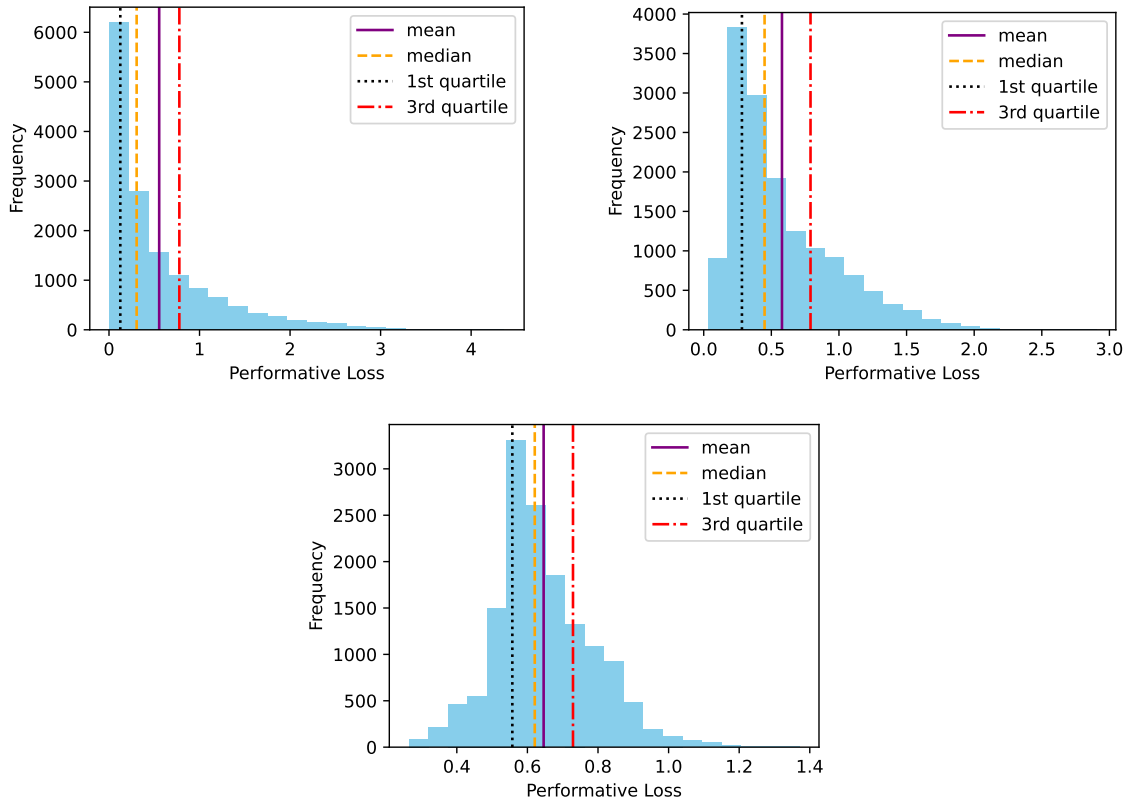

Figure IV.2: Histogram of performative loss under Experiment 4.5.1 with $\epsilon_{\text{true}} = 0.5$. Top left: histogram for the PO. Top right: histogram for the DRPO with $\rho = 0.02$. Bottom: hitogram for the DRPO with $\rho = 0.04$.

4.5.2 Partially Identifiable Distribution Map

Recall Example IV.2, we examine a location model for distribution map where the misspecification arises from the estimation error of the model parameter. Let $V = [\theta_1 - \theta_0 \mid \theta_2 - \theta_0 \mid \cdots \mid \theta_K - \theta_0] \in \mathbb{R}^{d\times K}$ and $U = [\mu_1 - \mu_0 \mid \mu_2 - \mu_0 \mid \cdots \mid \mu_K - \mu_0] \in \mathbb{R}^{d\times K}$ where $d = \dim(\Theta)$. The unknown parameter A can only be partially indentified through the equation $AV = U$ when $K < d$. A particular estimate of A is $\widehat{A} = UV^\dagger = U(V^\top V)^{-1}V^\top$, where $V^\dagger$ is the Moore-Penrose inverse of V. In fact, the parameter A is only identifiable up to the subspace $\mathcal{W} = \{UV^\dagger + E \mid \text{span}\{E^\top\} \subset \mathcal{N}(V^\top)\}$, where $\mathcal{N}(V^\top)$ is the left null space of V. Precicely, we have $AV = U$ if and only if $A \in \mathcal{W}$.

In this experiment, we still use the credit data. We observe sampling distribution of $\mathcal{D}(\mathbf{0})$, $\mathcal{D}(\mathbf{e}_1)$, and $\mathcal{D}(\mathbf{e}_2)$, where $\mathbf{e}_i$ is the i-th canonical basis. For the true distribution map, the performativity of the first two features is 0.5, while the performativity of the other 7 features is ϵ_{true}. In short, $A_{\text{true}} = \text{diag}(0.5, 0.5, \epsilon_{\text{true}}, \ldots, \epsilon_{\text{true}})$. We set the range of ϵ_{true} to be $[-0.5\eta, 0.5\eta]$ for $\eta \geq 0$. By using the estiamte $\widehat{A}$, we model the performativity of the first two features correctly, but wrongly model the other features as non-strategic. This time we fit TPO by Algorithm 3 instead of DRPO.

The top left plot of Figure IV.3 shows performative risk incurred by the PO and the TPO's with various tilt α's. As α increases, the TPO performs more uniformly across a wider range of ϵ_{true}. The top right plot of Figure IV.3 shows relative improvement in worst-case performative risk of the TPO to the PO as the tilt α increases, for different range of misspecification η's. Without misspeicification, $\eta = 0$, the PO is for sure better than the TPO. With moderate to large misspecification, $\eta \in \{2/3, 4/3, 2\}$, the TPO demonstrates certain advantages over the PO. The bottom plot of Figure IV.3 shows the relationship between the distributionally robust performative risk minimization and the tilted performative risk minimization: fitting DRPO with ρ (which returns the optimal dual variable $\mu^\star$) is equivalent to fitting TPO with $\alpha = 1/\mu^\star$.

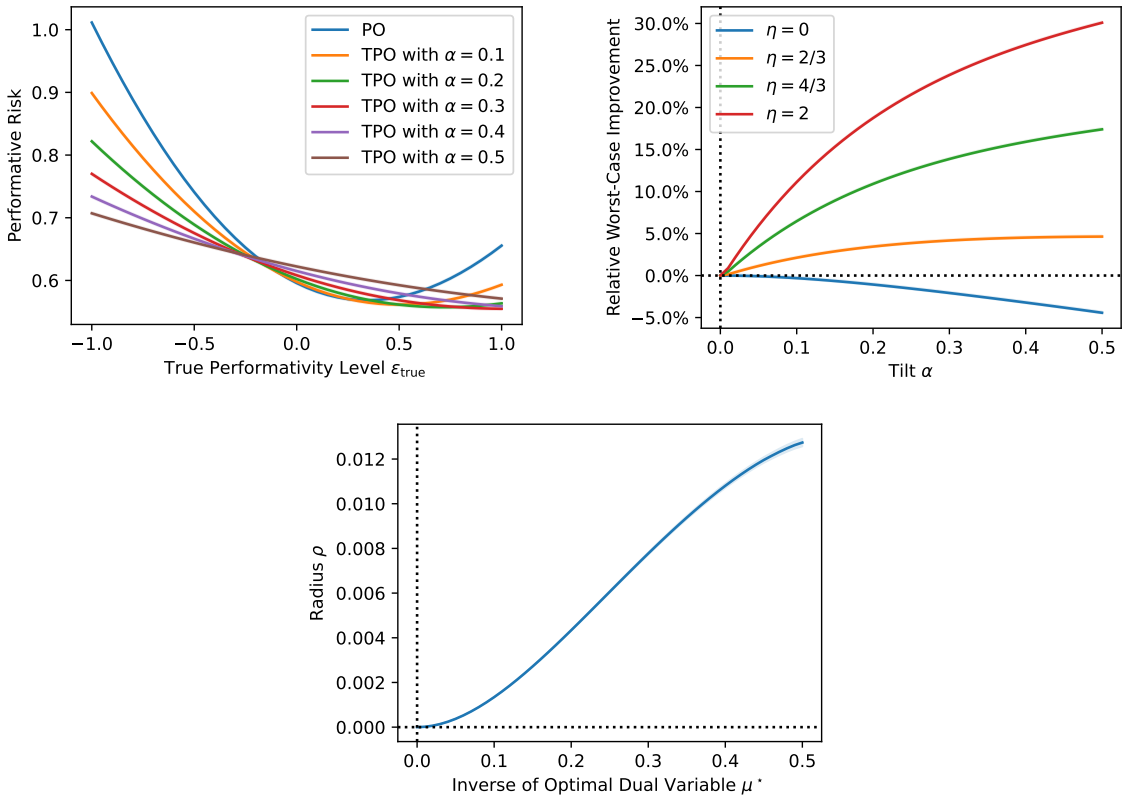

Figure IV.3: Results of Experiment 4.5.2. Top left: performative risk incurred by the PO and the TPO's with various tilt α's. Top right: relative improvement in worst-case performative risk of the TPO to the PO as the tilt α increases, for different range of misspecification η's. Bottom: the correspondence relationship between the radius ρ and the (inverse of) optimal dual variable $\mu^\star$.

4.5.3 Fairness without Demographics

Referencing to Example IV.3, we examine the scenario where the population distribution map is a mixture of two subpopulation distribution maps. We train a classifier using the population distribution map $\mathcal{D}_{\text{pop}}$, but target at its performance on both the majority and minority, $\mathcal{D}_{\text{maj}}$ and $\mathcal{D}_{\text{min}}$. The distribution map is therefore misspecified due to subpopulation shift.

The experimental setup resembles that in Peet-Pare et al. (2022). Note that the credit dataset used in the previous experiments lacks demographic features. To enable oracle access to demographic information, synthetic data is generated for a

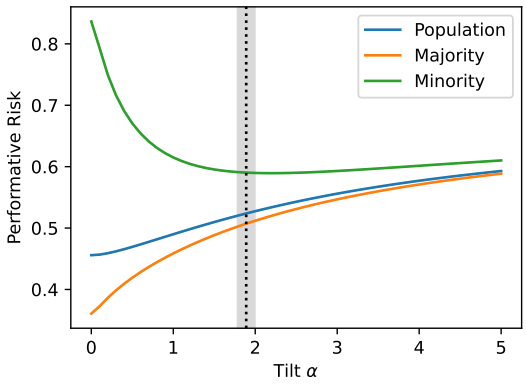

Figure IV.4: Results of Experiment 4.5.3. Performative risk of the population, the majority, and the minority, as the tilt α increases. The vertical band indicates the calibrated tilt α_{cal}'s.

performative classification task. The synthetic dataset exemplifies a scenario in which a linear decision boundary is not able to effectively classify both the majority and minority groups, necessitating a trade-off between them.

Figure IV.4 shows performative risk of the population, the majority, and the minority incurred by the TPO, as the tilt α increases. The PO (which is TPO with $\alpha = 0$) exhibits the lowest performative risk at the population, but the greatest disparity between its performance for the majority and minority groups. As α increases, the TPO reduces the performance gap between the two groups at the expense of an increased population performative risk. This suggests that the distributionally robust performative prediction framework has the potential to mitigate unfairness towards the minority group, even in the absence of demographic information. Using a small calibration set with demographics, we can calibrate the tilt α_{cal}'s via the calibration set approach described in Section 4.4.4. The goal is to calibrate the tilt to satisfy a four-fifth rule[2] with minimal population performative risk. The vertical bands in Figure IV.4 shows the calibrated tilt α_{cal}'s reasonably meet the goal.

[2]Performative risk for the minority is not 25% higher than that of the majority. This rule is motivated by the four-fifth rule documented in Uniform Guidelines on Employment Selection Procedures, 29 C.F.R. §1607.4(D) (2015).

4.6 Summary and Discussion

In this work, we present a distributionally robust performative prediction framework that aims to improve robustness against a variety of distribution maps, thereby mitigating the problem of distribution map misspecification. We show provable guarantees for DRPO as a robust approximation to the true PO when the nominal distribution map differs from the actual one. We developed efficient algorithms for minimizing the distributionally robust performative risk. Empirical experiments are conducted to support our theoretical findings and validate the proposed algorithms.

The extension of φ-divergence-based distributionally robust performative prediction to a Wasserstein DRO version is a natural direction for future research, calling for the development of new algorithms.

CHAPTER V

Concluding Remarks and Future Directions

This dissertation contributes to machine learning safety by addressing three concrete problems related to alignment, data quality, distribution robustness, and their intersections. In this concluding chapter, we summarize the contributions of the main chapters and suggest some future directions.

5.1 Calibrated Data-Dependent Constraints

In Chapter II, we consider the problem of exact constraint satisfaction probability in stochastic optimization with expected-value constraints. This problem has two components. First, a broad class of learning paradigms that involve data-dependent constraints can be formulated as a stochastic optimization problem with expected-value constraints, for example, learning with group fairness constraints. Second, the goal of this chapter is to improve constraint generalization in the sense of constraint satisfaction probability at the test time of the model's deployment. We propose a distributionally robust reformulation of data-dependent constraints to achieve this goal. Our theory guarantees that the proposed method satisfies the desired constraints with an asymptotically exact probability that the user specifies. On the algorithmic front, we design dual ascent algorithms to efficiently implement our method and its variants. The efficacy of our methods and the validity of our theory have been

demonstrated through extensive simulations and experiments.

Our method is implemented using the χ^2-divergence to facilitate theoretical analysis. We conjecture that our theory remains valid in the presence of a general φ-divergence. This is a promising future direction that necessitates the development of novel proof techniques. Additionally, it is worthwhile to investigate the application of our method to the inference of constrained M-estimation, a fundamental subject in mathematical statistics that has not been sufficiently studied in the context of expected-value constraints.

5.2 Statistical Cost of Data Noise in Constrained Learning

In Chapter III, we consider the problem of learning fair machine learning models with noisy sensitive attributes. The goal of this chapter is to study the statistical impact of privacy or missing data on fairness. We propose a statistically principled metric to quantify such impact. The proposed metric is a restricted notion of asymptotic statistical efficiency. Theoretically, we show that the statistical cost of noise on fairness is a relative decline in statistical efficiency. The relative efficiency loss is interpretable, as it explicitly depends on the privacy budget or the missing proportion, the subpopulation imbalance level, and a limited number of other problem-specific parameters. Our theory is substantiated by the experiments of fairness-sensitive classification tasks, which also demonstrate the usage of our tools.

The main result of this chapter has been applied to two specific scenarios of data noise: the locally differentially private sensitive attributes mechanism and the missing-at-random sensitive attributes mechanism. Naturally, it is of interest to study the interplay between fairness and data noise when the data is provided by other differential privacy mechanisms or is missing not at random.

5.3 Model Misspecification in Performative Prediction

In Chapter IV, we consider the problem of performative prediction for the interaction between a machine learning model and its environment. The goal of this chapter is to develop a robust learning framework that is more resilient to distribution map misspecification and that can more accurately approximate the true performative optimum. We propose a distributionally robust performative prediction framework to achieve this goal. Theoretically, we establish the strong duality, worst-case distribution map, and excess risk bounds for the solution concept of distributionally robust performative optimum (DRPO). Algorithmically, we design the distributionally robust performative risk minimization and the tilted performative risk minimization algorithms to find the DRPO efficiently. Furthermore, we offer practical advice regarding the calibration of the hyperparameters (the critical radius and the tilt). The advantage of our proposed methods is validated by extensive experiments that involve strategic classification, location family, and fairness without demographics.

It is possible to use Wasserstein distance to define the uncertainty collection within our framework. Our algorithms can be modified to compute Wasserstein DRPO. For example, one can establish the strong duality of Wasserstein DRPO and develop an alternating minimization algorithm similar to Algorithm 2. Nevertheless, the new algorithm involves an additional step of transport cost-regularized loss maximization due to the more complex dual form of Wasserstein DRO. For the corresponding theory, we expect that the square root of variance in (4.3.4) would be replaced by the Lipschitz norm of $\ell(\cdot, \theta_{\mathrm{PO,true}})$.

APPENDICES

APPENDIX A

Appendix of Chapter II

In Section A.1, A.2, A.3, we present proofs of theoretical results. In Section A.4, we summarize the dual ascent algorithm for solving (2.3.2). In Section A.5, we provide details for simulations on the multi-item newsvendor problem with independent constraints and run additional simulations. In Section A.6, we demonstrate how our method can be applied to ε-equality of opportunity. In Section A.7, we show the theoretical properties of the two-stage method for unknown active set. In Section A.8, we summarize the proxy dual ascent algorithm for handling non-differentiable constraints. In Section A.9, we provide details for Adult experiments. In Section A.10, we provide a standard derivation for the dual form of the robust constraint function (2.2.5). In Section A.11, experiments on additional baseline and dataset are conducted.

A.1 Proofs of Theorem 2.2.1 and Corollary 2.2.1

Note that Theorem 2.3.1 implies Theorem 2.2.1 and Corollary 2.3.1 implies Corollary 2.2.1 by letting $K = 1$. Therefore, it is sufficient to prove Theorem 2.3.1 and Corollary 2.3.1, whose proofs can be found in Appendix A.2 and A.3 respectively. $\square$

A.2 Proof of Theorem 2.3.1

Consider a stochastic optimization problem with K expected value constraints

$$(\mathcal{P}_0): \ \theta^\star \in \arg\min_{\theta\in\Theta} \left\{ \mathbb{E}_{Z\sim P_0}\left[f(\theta;Z)\right] : \mathbb{E}_{Z\sim P_0}\left[g_k(\theta;Z)\right] \le 0, k\in[K] \right\}.$$

Our proposed robust constraint method solves

$$(\mathcal{P}_n): \ \widehat{\theta}_{n,\boldsymbol{\rho}} \in \arg\min_{\theta\in\Theta} \left\{ \mathbb{E}_{Z\sim P_n}\left[f(\theta;Z)\right] : \sup_{D_\varphi(Q\|P_n)\le\rho_k/n} \mathbb{E}_{Z\sim Q}\left[g_k(\theta;Z)\right] \le 0, k\in[K] \right\},$$

where $\boldsymbol{\rho} = (\rho_1,\ldots,\rho_K)^\top$ is the collection of critical radii of uncertainty sets. Here we denote the empirical distribution $\widehat{P}_n$ by P_n for notation simplicity.

As a special case of our robust method, the sample average approximation (SAA) or empirical risk minimization (ERM) solves

$$\widehat{\theta}_{n,\mathbf{0}_K} \in \arg\min_{\theta\in\Theta} \left\{ \mathbb{E}_{Z\sim P_n}\left[f(\theta;Z)\right] : \mathbb{E}_{Z\sim P_n}\left[g_k(\theta;Z)\right] \le 0, k\in[K] \right\}.$$

We denote

$$F(\theta) = \mathbb{E}_{Z\sim P_0}\left[f(\theta;Z)\right] \quad \text{and} \quad \widehat{F}_n(\theta) = \mathbb{E}_{Z\sim P_n}\left[f(\theta;Z)\right],$$

$$G_k(\theta) = \mathbb{E}_{Z\sim P_n}\left[g_k(\theta;Z)\right] \quad \text{and} \quad \widehat{G}_{kn}(\theta) = \sup_{D_\varphi(Q\|P_n)\le\rho_k/n} \mathbb{E}_{Z\sim Q}\left[g_k(\theta;Z)\right] \text{ for } k\in[K],$$

and

$$\boldsymbol{G}(\theta) = \begin{pmatrix} G_1(\theta) \\ \vdots \\ G_K(\theta) \end{pmatrix} \quad \text{and} \quad \widehat{\boldsymbol{G}}_n(\theta) = \begin{pmatrix} \widehat{G}_{1n}(\theta) \\ \vdots \\ \widehat{G}_{Kn}(\theta) \end{pmatrix}.$$

Note that $\widehat{F}_n(\cdot)$ and $\widehat{G}_{kn}(\cdot)$'s are random functions serving as approximations to

$F(\cdot)$ and $G_k(\cdot)$'s. Consider the Lagrangian functions

$$L(\theta, \boldsymbol{\lambda}) = F(\theta) + \boldsymbol{\lambda}^\top \boldsymbol{G}(\theta) = F(\theta) + \sum_{k=1}^{K} \lambda_k G_k(\theta)$$

and

$$\widehat{L}_n(\theta, \boldsymbol{\lambda}) = \widehat{F}_n(\theta) + \boldsymbol{\lambda}^\top \widehat{\boldsymbol{G}}_n(\theta) = \widehat{F}_n(\theta) + \sum_{k=1}^{K} \lambda_k \widehat{G}_{kn}(\theta)$$

of the programs $(\mathcal{P}_0)$ and $(\mathcal{P}_n)$ respectively.

Lemma A.2.1 (Theorem 6.6.2 in Rubinstein and Shapiro (1993))**.** *Suppose that:*

(i) *The functions $F(\theta)$ and $G_k(\theta)$, $k \in [K]$, are twice continuously differentiable.*

(ii) *The true program $(\mathcal{P}_0)$ has a unique optimal solution $\theta^\star$ and a unique vector $\boldsymbol{\lambda}^\star$ of the Lagrange multipliers with $\theta^\star$ being an interior point of Θ.*

(iii) *The Hessian matrix $\nabla^2 L(\theta^\star, \boldsymbol{\lambda}^\star)$ is positive definite.*

(iv) *The random functions $\widehat{G}_{kn}(\theta)$,$k \in [K]$, are Lipschitz continuous in a neighborhood of $\theta^\star$ and differentiable at $\theta^\star$ with probability 1.*

(v)

$$\|\Delta_{in}(\theta^\star)\|_2 = O_p(n^{-1/2}), \quad i = 1, 2, 3$$

and there is a neighborhood U of $\theta^\star$ such that

$$\sup_{\theta \in U} \frac{\|\Delta_{in}(\theta) - \Delta_{in}(\theta^\star)\|_2}{n^{-1/2} + \|\theta - \theta^\star\|_2} = o_p(1), \quad i = 1, 2, 3.$$

Here we define random mappings $\Delta_{1n}(\theta) = \nabla \widehat{F}_n(\theta) - \nabla F(\theta)$, $\Delta_{2n}(\theta) = \widehat{\boldsymbol{G}}_n(\theta) - \boldsymbol{G}(\theta)$, and $\Delta_{3n}(\theta) = \nabla \widehat{\boldsymbol{G}}_n(\theta) - \nabla \boldsymbol{G}(\theta)$.

(vi) *Random vectors $\sqrt{n}(\nabla \widehat{L}_n(\theta^\star, \boldsymbol{\lambda}^\star), \widehat{\boldsymbol{G}}_n(\theta^\star))$ converge in distribution as $n \to \infty$ to a random vector $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2)$.*

Let $\widehat{\theta}_n$ be an optimal solution of $(\mathcal{P}_n)$ converging in probability as $n \to \infty$ to $\theta^\star$. Then

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{d} \bar{\boldsymbol{x}}(\boldsymbol{Y})$$

where $\bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}(\boldsymbol{Y})$ is the optimal solution to the quadratic programming problem

$$\begin{array}{ll} \underset{\boldsymbol{x}}{\text{minimize}} & \boldsymbol{x}^\top \boldsymbol{Y}_1 + \frac{1}{2}\boldsymbol{x}^\top \nabla^2 L(\theta^\star, \boldsymbol{\lambda}^\star)\boldsymbol{x} \\ \text{subject to} & \nabla \boldsymbol{G}(\theta^\star)^\top \boldsymbol{x} + \boldsymbol{Y}_2 = \boldsymbol{0} \end{array}.$$

From now on, we use $\varphi(t) = (t-1)^2$, which gives the χ^2-divergence. Lemma A.2.1 is adapted from Theorem 6.6.2 in Rubinstein and Shapiro (1993) under the strict complementarity assumption. Recall the standing assumptions, *(i)*, *(iv)*, *(v)* are guaranteed by the smoothness and concentration assumption, *(ii)* is postulated by the uniqueness assumption, and *(iii)* is ensured by the positive definiteness assumption. Now we derive the limiting distribution required in *(vi)*.

Let $\mathbb{B}$ denote the ℓ_2-ball of radius 1 in $\mathbb{R}^d$. According to Lemma 24 in Duchi and Namkoong (2019), for each $k \in [K]$ there exists $\epsilon_k > 0$ such that, with probability 1, there exists an N_k such that $n \geq N_k$ implies

$$\widehat{G}_{kn}(\theta) = \mathbb{E}_{P_n}[g_k(\theta; Z)] + \sqrt{\frac{\rho_k \operatorname{Var}_{P_n}[g_k(\theta; Z)]}{n}} \text{ for all } \theta \in \theta^\star + \epsilon_k \mathbb{B}.$$

Taking $\epsilon_0 = \min\{\epsilon_k : k \in [K]\}$ and $N_0 = \max\{N_k : k \in [K]\}$, we have the following uniform expansion holds, that is,

$$\widehat{G}_{kn}(\theta) = \mathbb{E}_{P_n}[g_k(\theta; Z)] + \sqrt{\frac{\rho_k}{n} \operatorname{Var}_{P_n}[g_k(\theta; Z)]} \text{ for all } k \in [K] \text{ and } \theta \in \theta^\star + \epsilon_0 \mathbb{B}$$

P_0-almost surely given $n \geq N_0$.

Therefore, for sufficiently large n, for $k \in [K]$ we have

$$
\begin{aligned}
&\widehat{G}_{kn}(\theta^\star) \\
&= \mathbb{E}_{P_n}[g_k(\theta^\star; Z)] + \sqrt{\frac{\rho_k}{n} \operatorname{Var}_{P_n}[g_k(\theta; Z)]} \\
&= \mathbb{E}_P[g_k(\theta^\star; Z)] + \{\mathbb{E}_{P_n}[g_k(\theta^\star; Z)] - \mathbb{E}_{P_0}[g_k(\theta^\star; Z)]\} + \sqrt{\frac{\rho_k}{n} \left(\operatorname{Var}_{P_0}[g_k(\theta^\star; Z)] + o_P(1)\right)} \\
&= G_k(\theta^\star) + \{\mathbb{E}_{P_n}[g_k(\theta^\star; Z)] - \mathbb{E}_{P_0}[g_k(\theta^\star; Z)]\} + \sqrt{\frac{\rho_k}{n} \operatorname{Var}_{P_0}[g_k(\theta^\star; Z)]} + o_P(n^{-1/2})
\end{aligned}
$$

and

$$
\begin{aligned}
\nabla \widehat{G}_{kn}(\theta^\star) &= \mathbb{E}_{P_n}[\nabla g_k(\theta^\star; Z)] + \nabla \sqrt{\frac{\rho_k}{n} \operatorname{Var}_{P_n}[g_k(\theta; Z)]} \\
&= \mathbb{E}_P[\nabla g_k(\theta^\star; Z)] + \{\mathbb{E}_{P_n}[\nabla g_k(\theta^\star; Z)] - \mathbb{E}_{P_0}[\nabla g_k(\theta^\star; Z)]\} + \\
&\quad \sqrt{\frac{\rho_k}{n}} \frac{\mathbb{E}_{P_n}[(\nabla g_k(\theta^\star, X) - \mathbb{E}_{P_n}[\nabla g_k(\theta^\star; Z)])\,(g_k(\theta^\star; Z) - \mathbb{E}_{P_n}[g_k(\theta^\star; Z)])]}{\sqrt{\operatorname{Var}_{P_n}[g_k(\theta^\star; Z)]}} \\
&= \nabla \mathbb{E}_{P_0}[g_k(\theta^\star; Z)] + \{\mathbb{E}_{P_n}[\nabla g_k(\theta^\star; Z)] - \mathbb{E}_{P_0}[\nabla g_k(\theta^\star; Z)]\} + \\
&\quad \sqrt{\frac{\rho_k}{n}} \left(\frac{\operatorname{Cov}_{P_0}(\nabla g_k(\theta^\star; Z), g_k(\theta^\star; Z))}{\sqrt{\operatorname{Var}_{P_0}[g_k(\theta^\star; Z)]}} + o_P(1) \right) \\
&= \nabla G_k(\theta^\star) + \{\mathbb{E}_{P_n}[\nabla g_k(\theta^\star; Z)] - \mathbb{E}_{P_0}[\nabla g_k(\theta^\star; Z)]\} + \\
&\quad \sqrt{\frac{\rho_k}{n}} \frac{\operatorname{Cov}_{P_0}(\nabla g_k(\theta^\star; Z), g_k(\theta^\star; Z))}{\sqrt{\operatorname{Var}_{P_0}[g_k(\theta^\star; Z)]}} + o_P(n^{-1/2}).
\end{aligned}
$$

For the objective function and its empirical counterpart, we have

$$
\widehat{F}(\theta^\star) = \mathbb{E}_{P_n}[f(\theta^\star; Z)] = F(\theta^\star) + \{\mathbb{E}_{P_n}[f(\theta^\star; Z)] - \mathbb{E}_{P_0}[f(\theta^\star; Z)]\}
$$

and

$$
\nabla \widehat{F}(\theta^\star) = \mathbb{E}_{P_n}[\nabla f(\theta^\star; Z)] = \nabla F(\theta^\star) + \{\mathbb{E}_{P_n}[\nabla f(\theta^\star; Z)] - \mathbb{E}_{P_0}[\nabla f(\theta^\star; Z)]\}.
$$

We derive the limiting distribution of random vectors $\sqrt{n}(\nabla \widehat{L}_n(\theta^\star, \boldsymbol{\lambda}^\star), \widehat{\boldsymbol{G}}_n(\theta^\star))$. For simplicity of notations, we denote $P_n m = \mathbb{E}_{P_n}[m(\theta^\star; Z)]$ and $P_0 m = \mathbb{E}_{P_0}[m(\theta^\star; Z)]$ for any random function $m(\theta; Z)$, $\mathrm{Cov}(\nabla g_k, g_k) = \mathrm{Cov}_{P_0}(\nabla g_k(\theta^\star; Z), g_k(\theta^\star; Z))$ and $\mathrm{Var}[g_k] = \mathrm{Var}_{P_0}[g_k(\theta^\star; Z)]$. Then we have

$$
\begin{aligned}
&\begin{bmatrix} \nabla \widehat{L}_n(\theta^\star, \boldsymbol{\lambda}^\star) \\ \widehat{\boldsymbol{G}}_n(\theta^\star) \end{bmatrix} \\
&= \begin{bmatrix} \nabla \widehat{F}_n(\theta^\star) + \sum_{k=1}^K \lambda_k^\star \nabla \widehat{G}_{kn}(\theta^\star) \\ \vdots \\ \widehat{G}_{kn}(\theta^\star) \\ \vdots \end{bmatrix} \\
&= \begin{bmatrix} \mathrm{Term}_1 + (P_n \nabla f - P_0 \nabla f) + \sum_{k=1}^K \lambda_k^\star (P_n \nabla g_k - P_0 \nabla g_k) + \mathrm{Term}_2 \\ \vdots \\ \underbrace{G_k(\theta^\star)}_{=0 \text{ due to active constraint}} + (P_n g_k - P_0 g_k) + \sqrt{\frac{\rho_k}{n} \mathrm{Var}[g_k]} \\ \vdots \end{bmatrix} \\
&\quad + o_P\left(\frac{1}{\sqrt{n}}\right) \\
&= \frac{1}{\sqrt{n}} \begin{bmatrix} \sum_{k=1}^K \frac{\lambda_k^\star \sqrt{\rho_k}\, \mathrm{Cov}(\nabla g_k, g_k)}{\sqrt{\mathrm{Var}[g_k]}} \\ \vdots \\ \sqrt{\rho_k \, \mathrm{Var}[g_k]} \\ \vdots \end{bmatrix} + \\
&\quad \left\{ P_n \begin{bmatrix} \nabla f + \sum_{k=1}^K \lambda_k^\star \nabla g_k \\ \vdots \\ g_k \\ \vdots \end{bmatrix} - P_0 \begin{bmatrix} \nabla f + \sum_{k=1}^K \lambda_k^\star \nabla g_k \\ \vdots \\ g_k \\ \vdots \end{bmatrix} \right\} + o_P\left(\frac{1}{\sqrt{n}}\right),
\end{aligned}
$$

where

$$\text{Term}_1 := \nabla F(\theta^\star) + \sum_{k=1}^{K} \lambda_k^\star \nabla G_k(\theta^\star)$$

and $\text{Term}_1 = 0$ due to KKT condition, and

$$\text{Term}_2 := \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \frac{\lambda_k^\star \sqrt{\rho_k}\, \text{Cov}(\nabla g_k, g_k)}{\sqrt{\text{Var}[g_k]}}.$$

By central limit theorem,

$$\sqrt{n} \left\{ P_n \begin{bmatrix} \nabla f + \sum_{k=1}^K \lambda_k^\star \nabla g_k \\ \vdots \\ g_k \\ \vdots \end{bmatrix} - P_0 \begin{bmatrix} \nabla f + \sum_{k=1}^K \lambda_k^\star \nabla g_k \\ \vdots \\ g_k \\ \vdots \end{bmatrix} \right\}$$
$$\xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right),$$

where

$$\Sigma_{11} = \text{Var}_{P_0}\left[\nabla f(\theta^\star; Z) + \sum_{k=1}^{K} \lambda_k^\star \nabla g_k(\theta^\star; Z) \right] \in \mathbb{R}^{d \times d},$$

$$\Sigma_{12} = \text{Cov}_{P_0}\left(\nabla f(\theta^\star; Z) + \sum_{k=1}^{K} \lambda_k^\star \nabla g_k(\theta^\star; Z), \boldsymbol{G}(\theta^\star; Z) \right) \in \mathbb{R}^{d \times K},$$

$$\Sigma_{21} = \Sigma_{12}^\top,$$

$$\Sigma_{22} = \text{Var}_{P_0}[\boldsymbol{G}(\theta^\star; Z)] \in \mathbb{R}^{K \times K}.$$

By Slutsky's theorem,

$$
\begin{aligned}
&\sqrt{n}\begin{bmatrix}\nabla\widehat{L}_n(\theta^\star,\boldsymbol{\lambda}^\star)\\ \widehat{\boldsymbol{G}}_n(\theta^\star)\end{bmatrix}\\
&=\begin{bmatrix}\sum_{k=1}^K \frac{\lambda_k^\star\sqrt{\rho_k}\,\mathrm{Cov}(\nabla g_k, g_k)}{\sqrt{\mathrm{Var}[g_k]}}\\ \vdots\\ \sqrt{\rho_k\,\mathrm{Var}[g_k]}\\ \vdots\end{bmatrix}+\\
&\sqrt{n}\left\{P_n\begin{bmatrix}\nabla f+\sum_{k=1}^K\lambda_k^\star\nabla g_k\\ \vdots\\ g_k\\ \vdots\end{bmatrix}-P_0\begin{bmatrix}\nabla f+\sum_{k=1}^K\lambda_k^\star\nabla g_k\\ \vdots\\ g_k\\ \vdots\end{bmatrix}\right\}+o_P(1)\\
&\xrightarrow{d}\mathcal{N}\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix},\begin{bmatrix}\Sigma_{11}&\Sigma_{12}\\ \Sigma_{21}&\Sigma_{22}\end{bmatrix}\right),
\end{aligned}
$$

where

$$
\begin{aligned}
\mu_1&=\sum_{k=1}^K\frac{\lambda_k^\star\sqrt{\rho_k}\,\mathrm{Cov}_{P_0}(\nabla g_k(\theta^\star;Z),g_k(\theta^\star;Z))}{\sqrt{\mathrm{Var}_{P_0}[g_k(\theta^\star;Z)]}}\in\mathbb{R}^d,\\
\mu_2&=\begin{bmatrix}\sqrt{\rho_1\,\mathrm{Var}_{P_0}[g_1(\theta^\star;Z)]}\\ \vdots\\ \sqrt{\rho_K\,\mathrm{Var}_{P_0}[g_K(\theta^\star;Z)]}\end{bmatrix}\in\mathbb{R}^K.
\end{aligned}
$$

Therefore, we conclude that the limiting distribution of $\sqrt{n}(\nabla\widehat{L}_n(\theta^\star,\boldsymbol{\lambda}^\star),\widehat{\boldsymbol{G}}_n(\theta^\star))$ is

$$
(\boldsymbol{Y}_1,\boldsymbol{Y}_2)\sim\mathcal{N}\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix},\begin{bmatrix}\Sigma_{11}&\Sigma_{12}\\ \Sigma_{21}&\Sigma_{22}\end{bmatrix}\right).
$$

By Lemma (A.2.1), we have

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{d} \bar{\boldsymbol{x}},$$

where $\bar{\boldsymbol{x}}$ is given by the linear system

$$\underbrace{\begin{bmatrix} \nabla^2 L(\theta^\star, \boldsymbol{\lambda}^\star) & \nabla \boldsymbol{G}(\theta^\star) \\ \nabla \boldsymbol{G}(\theta^\star)^\top & \mathbf{0} \end{bmatrix}}_{\triangleq B} \begin{bmatrix} \bar{\boldsymbol{x}} \\ \bar{\boldsymbol{\lambda}} \end{bmatrix} = - \begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{bmatrix} \sim \mathcal{N}\left(- \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right),$$

or

$$\begin{bmatrix} \bar{\boldsymbol{x}} \\ \bar{\boldsymbol{\lambda}} \end{bmatrix} \sim \mathcal{N}\left(-B^{-1} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, B^{-1} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} B^{-1} \right), \tag{A.2.1}$$

which implies $\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{d} \bar{\boldsymbol{x}} \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$ for some $\bar{\mu}$ and $\bar{\Sigma}$ determined by (A.2.1).

By delta method, we have

$$\begin{aligned} &\sqrt{n} \begin{bmatrix} \mathbb{E}_{P_0}[g_1(\widehat{\theta}_n; Z)] \\ \vdots \\ \mathbb{E}_{P_0}[g_K(\widehat{\theta}_n; Z)] \end{bmatrix} \\ &= \sqrt{n}\boldsymbol{G}(\widehat{\theta}_n) = \sqrt{n}\{\boldsymbol{G}(\widehat{\theta}_n) - \underbrace{\boldsymbol{G}(\theta^\star)}_{=0}\} \xrightarrow{d} \mathcal{N}(\nabla \boldsymbol{G}(\theta^\star)^\top \bar{\mu}, \nabla \boldsymbol{G}(\theta^\star)^\top \bar{\Sigma} \nabla \boldsymbol{G}(\theta^\star)). \end{aligned}$$

Now we calculate $\nabla \boldsymbol{G}(\theta^\star)^\top \bar{\mu}$ and $\nabla \boldsymbol{G}(\theta^\star)^\top \bar{\Sigma} \nabla \boldsymbol{G}(\theta^\star))$.

For notation simplicity, we denote $\nabla^2 L = \nabla^2 L(\theta^\star, \boldsymbol{\lambda}^\star), \nabla \boldsymbol{G} = \nabla \boldsymbol{G}(\theta^\star)$ and $H = (\nabla^2 L)^{-1} \nabla \boldsymbol{G} [\nabla \boldsymbol{G}^\top (\nabla^2 L)^{-1} \nabla \boldsymbol{G}]^{-1}$. By block matrix inversion, we have

$$B^{-1} = \begin{bmatrix} (\nabla^2 L)^{-1} - H \nabla \boldsymbol{G}^\top (\nabla^2 L)^{-1} & H \\ H^\top & -[\nabla \boldsymbol{G}^\top (\nabla^2 L)^{-1} \nabla \boldsymbol{G}]^{-1}. \end{bmatrix}$$

By (A.2.1), we have

$$\bar{\mu} = -\left\{(\nabla^2 L)^{-1} - H\nabla\boldsymbol{G}^\top(\nabla^2 L)^{-1}\right\}\mu_1 - H\mu_2.$$

Note that $\nabla\boldsymbol{G}^\top H = \boldsymbol{I}_K$ and $\nabla\boldsymbol{G}^\top\left\{(\nabla^2 L)^{-1} - H\nabla\boldsymbol{G}^\top(\nabla^2 L)^{-1}\right\} = \mathbf{0}_{K\times K}$. We have

$$\nabla\boldsymbol{G}(\theta^\star)^\top\bar{\mu} = -\nabla\boldsymbol{G}^\top\left\{(\nabla^2 L)^{-1} - H\nabla\boldsymbol{G}^\top(\nabla^2 L)^{-1}\right\}\mu_1 - \nabla\boldsymbol{G}H\mu_2 = -\mu_2$$

and

$$\begin{aligned}
&\nabla\boldsymbol{G}(\theta^\star)^\top\bar{\Sigma}\nabla\boldsymbol{G}(\theta^\star)\\
&= \nabla\boldsymbol{G}^\top\left[\left\{(\nabla^2 L)^{-1} - H\nabla\boldsymbol{G}^\top(\nabla^2 L)^{-1}\right\}\Sigma_{11} + H\Sigma_{21}\right]\\
&\quad\underbrace{\left\{(\nabla^2 L)^{-1} - (\nabla^2 L)^{-1}\nabla\boldsymbol{G}H^\top\right\}\nabla\boldsymbol{G}}_{=\mathbf{0}_{K\times K}}\\
&\quad+\underbrace{\nabla\boldsymbol{G}^\top\left\{(\nabla^2 L)^{-1} - H\nabla\boldsymbol{G}^\top(\nabla^2 L)^{-1}\right\}}_{=\mathbf{0}_{K\times K}}\Sigma_{12}H^\top\nabla\boldsymbol{G} + \nabla\boldsymbol{G}^\top H\Sigma_{22}H^\top\nabla\boldsymbol{G}\\
&= \Sigma_{22}.
\end{aligned}$$

Therefore, we conclude that

$$\begin{aligned}
\sqrt{n}\begin{bmatrix}\mathbb{E}_{P_0}[g_1(\widehat{\theta}_n; Z)]\\ \vdots \\ \mathbb{E}_{P_0}[g_K(\widehat{\theta}_n; Z)]\end{bmatrix} &\xrightarrow{d} \mathcal{N}(-\mu_2, \Sigma_{22})\\
&\overset{d}{=} \mathcal{N}\left(-\begin{bmatrix}\sqrt{\rho_1\operatorname{Var}_{P_0}[g_1(\theta^\star; Z)]}\\ \vdots \\ \sqrt{\rho_K\operatorname{Var}_{P_0}[g_K(\theta^\star; Z)]}\end{bmatrix}, \operatorname{Var}_{P_0}\begin{bmatrix}g_1(\theta^\star; Z)\\ \vdots \\ g_K(\theta^\star; Z)\end{bmatrix}\right).
\end{aligned}$$

Hence we complete the proof of Theorem 2.3.1. □

A.3 Proof of Corollary 2.3.1

Recall that $D = \mathrm{diag}(\mathrm{Var}_{P_0}[g_1(\theta^\star; Z)], \ldots, \mathrm{Var}_{P_0}[g_K(\theta^\star; Z)])$. According to Theorem 2.3.1, we have

$$
\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}\left\{\widehat{\theta}_n \text{ is feasible}\right\} \\
&= \lim_{n\to\infty} \mathbb{P}\left\{ \begin{bmatrix} \mathbb{E}_{P_0}[g_1(\widehat{\theta}_n; Z)] \\ \vdots \\ \mathbb{E}_{P_0}[g_K(\widehat{\theta}_n; Z)] \end{bmatrix} \in -\mathbb{R}_+^K \right\} \\
&= \lim_{n\to\infty} \mathbb{P}\left\{ \sqrt{n} \begin{bmatrix} \mathbb{E}_{P_0}[g_1(\widehat{\theta}_n; Z)] \\ \vdots \\ \mathbb{E}_{P_0}[g_K(\widehat{\theta}_n; Z)] \end{bmatrix} \leq \mathbf{0}_K \right\} \\
&= \mathbb{P}\left\{ \mathcal{N}\left(- \begin{bmatrix} (\rho_1 \mathrm{Var}_{P_0}[g_1(\theta^\star; Z)])^{\frac{1}{2}} \\ \vdots \\ (\rho_K \mathrm{Var}_{P_0}[g_K(\theta^\star; Z)])^{\frac{1}{2}} \end{bmatrix}, \mathrm{Var}_{P_0} \begin{bmatrix} g_1(\theta^\star; Z) \\ \vdots \\ g_K(\theta^\star; Z) \end{bmatrix} \right) \leq \mathbf{0}_K \right\} \\
&= \mathbb{P}\left\{ \mathcal{N}\left(-D^{-\frac{1}{2}} \begin{bmatrix} (\rho_1 \mathrm{Var}_{P_0}[g_1(\theta^\star; Z)])^{\frac{1}{2}} \\ \vdots \\ (\rho_K \mathrm{Var}_{P_0}[g_K(\theta^\star; Z)])^{\frac{1}{2}} \end{bmatrix}, D^{-\frac{1}{2}} \mathrm{Var} \begin{bmatrix} g_1(\theta^\star; Z) \\ \vdots \\ g_K(\theta^\star; Z) \end{bmatrix} D^{-\frac{1}{2}} \right) \leq D^{-\frac{1}{2}} \mathbf{0}_K \right\} \\
&= \mathbb{P}\left\{ \mathcal{N}\left(- \begin{bmatrix} \sqrt{\rho_1} \\ \vdots \\ \sqrt{\rho_K} \end{bmatrix}, \mathrm{Corr}_{P_0} \begin{bmatrix} g_1(\theta^\star; Z) \\ \vdots \\ g_K(\theta^\star; Z) \end{bmatrix} \right) \leq \mathbf{0}_K \right\} \\
&= \mathbb{P}\left\{ \mathcal{N}\left(\mathbf{0}_K, \mathrm{Corr}_{P_0} \begin{bmatrix} g_1(\theta^\star; Z) \\ \vdots \\ g_K(\theta^\star; Z) \end{bmatrix} \right) \leq \begin{bmatrix} \sqrt{\rho_1} \\ \vdots \\ \sqrt{\rho_K} \end{bmatrix} \right\}.
\end{aligned}
$$

Hence we complete the proof of Corollary 2.3.1. □

A.4 Dual Ascent Algorithm for (2.3.2)

We summarize the dual ascent algorithm for solving (2.3.2) in Algorithm 4. Similar to Algorithm 1, the main cost of Algorithm 4 is incurred in the evaluation of dual function. The dual function evaluation step is still suitable for stochastic approximation. Therefore, the total cost of Algorithm 4 is comparable to that of a standard distributionally robust optimization problem.

Algorithm 4 Dual ascent algorithm for (2.3.2)

1: **Input:** starting dual iterate $\boldsymbol{\lambda}_0 = (\lambda_{01}, \ldots, \lambda_{0K})^\top \in \mathbb{R}_+^K$
2: **repeat**
3: Evaluate dual function:

$$(\theta_t, \boldsymbol{\mu}_t, \boldsymbol{\nu}_t) \leftarrow \arg\min_{\theta, \boldsymbol{\mu} \in \mathbb{R}_+^K, \boldsymbol{\nu}} \Big(\tfrac{1}{n} \textstyle\sum_{i=1}^n f(\theta; Z_i)$$

$$+ \textstyle\sum_{k=1}^K \lambda_{tk} \left\{ \tfrac{1}{n} \sum_{i=1}^n \mu_{tk} \varphi^*\big(\tfrac{g_k(\theta; Z_i) - \nu_{tk}}{\mu_{tk}}\big) + \mu_{tk} \tfrac{\rho_k}{n} + \nu_{tk} \right\} \Big)$$

4: Dual ascent update:

$$\lambda_{t+1,k} \leftarrow \left[\lambda_{tk} + \eta_t \left\{ \tfrac{1}{n} \textstyle\sum_{i=1}^n \mu_{tk} \varphi^*\big(\tfrac{g_k(\theta_t; Z_i) - \nu_{tk}}{\mu_{tk}}\big) + \mu_{tk} \tfrac{\rho_\alpha}{n} + \nu_{tk} \right\} \right]_+, k \in [K]$$

5: **until** converged

A.5 Simulations: Details and More

In this section, we provide details for simulations on the multi-item newsvendor problem with independent constraints (which we present in the main text), and conduct more simulations on: (1) multi-item newsvendor problem with dependent constraints, and (2) single-item newsvendor problem with a single constraint.

A.5.1 Multi-item Newsvendor Problem

First recall that in Section 2.4, we simulate the frequency of constraint satisfaction for the following multi-item newsvendor problem:

$$\begin{aligned} \max_{\theta\in\Theta} \quad & \mathbb{E}_{P_0}\left[p^\top \min\{Z,\theta\} - c^\top\theta\right] \\ \text{subject to} \quad & \mathbb{E}_{P_0}[(\|Z^{(1)}\|_2^2 - \|\theta^{(1)}\|_2^2)_+] \le \varepsilon_1 \\ & \mathbb{E}_{P_0}[(\|Z^{(2)}\|_2^2 - \|\theta^{(2)}\|_2^2)_+] \le \varepsilon_2 \end{aligned} \tag{A.5.1}$$

where $c \in \mathbb{R}^d_+$ is the manufacturing cost, $p \in \mathbb{R}^d_+$ is the sell price, $\theta \in \Theta = [0,100]^d$ is the number of items in stock, $Z \in \mathbb{R}^d$ is a random variable with probability distribution P_0 representing the demand, and there are d items in total. The distribution P_0 is unknown but we observe IID samples $Z_1,\ldots,Z_n$ from P_0. All of the items have been partitioned into two groups so that the corresponding demand and stock can be written as $Z = (Z^{(1)}, Z^{(2)})$ and $\theta = (\theta^{(1)}, \theta^{(2)})$. The constraints in the problem exclude stock levels that underestimate the demand too much for each group of items, where $\varepsilon_1, \varepsilon_2 > 0$ indicate tolerance level of such underestimation. The target of the problem is to maximize the profit while satisfying the constraints.

We can rewrite the maximization problem (A.5.1) as a minimization problem with expected value constraints in the form of (2.3.1), that is,

$$\begin{aligned} \min_{\theta\in\Theta} \quad & \mathbb{E}_{P_0}\left[c^\top\theta - p^\top \min\{Z,\theta\}\right] \\ \text{subject to} \quad & \mathbb{E}_{P_0}[(\|Z^{(1)}\|_2^2 - \|\theta^{(1)}\|_2^2)_+ - \varepsilon_1] \le 0 \\ & \mathbb{E}_{P_0}[(\|Z^{(2)}\|_2^2 - \|\theta^{(2)}\|_2^2)_+ - \varepsilon_2] \le 0 \end{aligned} \tag{A.5.2}$$

so that we can apply our method (2.3.2).

We generate $Z_1,\ldots,Z_n$ IID from multivariate normal distribution $\mathcal{N}(\mu,\Sigma)$. In addition, we set the number of items $d = 4$, the mean of the normal distribution $\mu = (10,10,10,10)^\top$, the cost $c = (1,1,1,1)^\top$, the price $p = (2,2,2,2)^\top$, and the

tolerance level of underestimation $(\varepsilon_1, \varepsilon_2) = (1, 1)$. Moreover, we partition the items into the group of the first two items and the group of the last two items. We solve the empirical problem with robust constraint:

$$\begin{array}{ll} \min_{\theta\in\Theta} & \mathbb{E}_{\widehat{P}_n}\left[c^\top\theta - p^\top \min\{Z, \theta\}\right] \\ \text{subject to} & \sup_{D_\varphi(P\|\widehat{P}_n)\le\rho/n} \mathbb{E}_P[(\|Z^{(1)}\|_2^2 - \|\theta^{(1)}\|_2^2)_+ - \varepsilon_1] \le 0 \\ & \sup_{D_\varphi(P\|\widehat{P}_n)\le\rho/n} \mathbb{E}_P[(\|Z^{(2)}\|_2^2 - \|\theta^{(2)}\|_2^2)_+ - \varepsilon_2] \le 0 \end{array} \tag{A.5.3}$$

using samples of size $n \in \{30, 300, 3000\}$.

For the covariance of the multivariate normal distribution, we consider a exchangeable correlation structure

$$\Sigma = 9 \times \begin{bmatrix} 1 & r & r & r \\ r & 1 & r & r \\ r & r & 1 & r \\ r & r & r & 1 \end{bmatrix}$$

We solve (A.5.3) with $\boldsymbol{\rho} = (\rho, \rho)^\top = (z_\alpha, z_\alpha)^\top$ for $\alpha \in \{0.4, 0.25, 0.1, 0.05, 0.005\}$, of which the corresponding $\sqrt{\rho} \in \{0.253, 0.674, 1.281, 1.644, 2.575\}$.

Independent constraints In Section 2.4, we consider $r = 0$ so that the two constraints are generally uncorrelated with each other. As suggested by the asymptotic theory in Section 2.3, the nominal probability of constraint satisfaction is $1 - \alpha$ for each constraint and $(1 - \alpha)^2$ for both constraints due to the independence of two constraints. The results are discussed in Section 2.4.

Dependent constraints Now we consider $r = 0.6$ so that the two constraints are generally correlated with each other. As suggested by the asymptotic theory in Section 2.3, the nominal probability of constraint satisfaction is $1 - \alpha$ for each constraint. The nominal probability of constraint satisfaction for both constraints is no longer $(1 - \alpha)^2$ due to the dependence of two constraints, but the probability is

given by (here $r^\star = \mathrm{Corr}_{P_0}(g_1(\theta^\star; Z), g_2(\theta^\star; Z))$ in the following equation)

$$\mathbb{P}\left\{\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & r^\star\\ r^\star & 1\end{bmatrix}\right) \leq_{\mathbb{R}^2} \begin{bmatrix}\sqrt{\rho}\\ \sqrt{\rho}\end{bmatrix}\right\}.$$

In Figure A.1, we plot frequencies of constraint satisfaction for each constraint and both constraints, all of which are averaged over 1000 replicates. As the sample size n grows, the frequency versus probability curve converges to the theoretical dashed line of limiting probability of constraint satisfaction, validating our theory in the large sample regime. We note that in this example, the frequency of constraint satisfaction is higher than that of the experiments with independent constraints, for each $\boldsymbol{\rho}$.

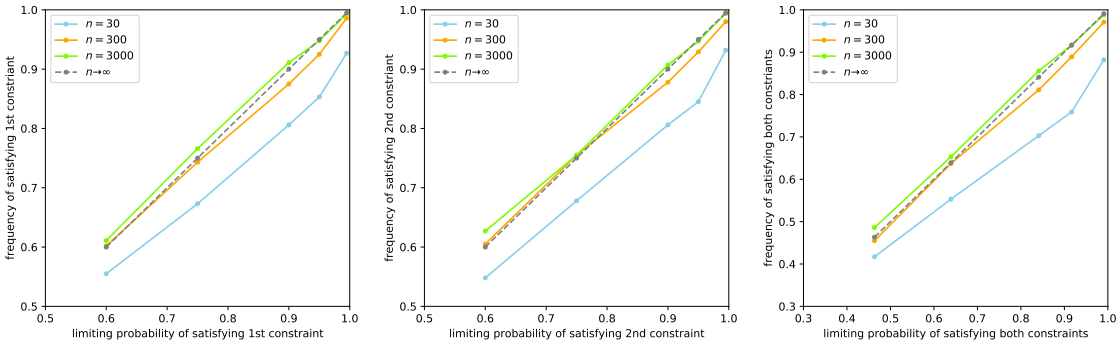

Figure A.1: Frequency versus limiting probability of constraint satisfaction of the first constraint (left), the second constraint (middle), and both of the constraints (right).

A.5.2 Single-item Newsvendor Problem

In this subsection, we consider the following single-item newsvendor problem:

$$\begin{aligned} \max_{\theta\in\Theta} \quad & \mathbb{E}_{P_0}\left[p\min\{Z,\theta\} - c\theta\right] \\ \text{subject to} \quad & \mathbb{E}_{P_0}\left[(Z-\theta)_+\right] \leq \varepsilon \end{aligned}$$

where $c > 0$ is the manufacturing cost, $p \geq c$ is the sell price, $\theta \in \Theta = [0, 100]$ is the number of items in stock, and Z is a random variable with probability distribution P_0 representing the demand. The distribution P_0 is unknown but instead we observe

IID samples $Z_1, \ldots, Z_n$ from P. The constraint in the problem excludes stocking levels that underestimate the demand too much, where $\varepsilon > 0$ indicates tolerance level of such underestimation. The target of the problem is to maximize the profit while satisfying the constraint. Note that the problem is equivalent to

$$\begin{aligned} \min_{\theta \in \Theta} \quad & \mathbb{E}_{P_0}\big[c\theta - p\min\{Z, \theta\}\big] \\ \text{subject to} \quad & \mathbb{E}_{P_0}\big[(Z-\theta)_+ - \varepsilon\big] \leq 0 \end{aligned}$$

which is a particular case of (2.1.1).

We generate $Z_1, \ldots, Z_n$ IID from exponential distribution with mean 10. In addition, we set the cost $c = 1$, the price $p = 2$, and the tolerance level of underestimation $\varepsilon = 1$. We solve the empirical problem with robust constraint:

$$\begin{aligned} \min_{\theta \in \Theta} \quad & \mathbb{E}_{\widehat{P}_n}\big[c\theta - p\min\{Z, \theta\}\big] \\ \text{subject to} \quad & \sup_{D_\varphi(P\|\widehat{P}_n) \leq \rho/n} \mathbb{E}_P\big[(Z-\theta)_+ - \varepsilon\big] \leq 0 \end{aligned}$$

using samples of size $n \in \{300, 3000, 30000\}$.

In Figure A.2, we plot frequencies of constraint satisfaction which are averaged over 1000 replicates. As the sample size n grows, the frequency versus probability curve converges to the theoretical dashed line of limiting probability of constraint satisfaction, validating our theory in the large sample regime.

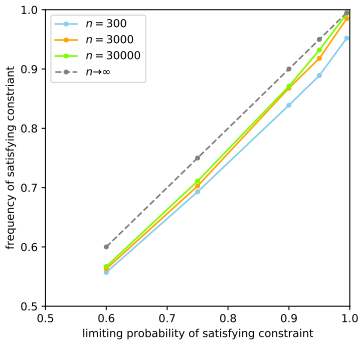

Figure A.2: Frequency versus limiting probability of constraint satisfaction.

A.6 Application to ε-Equality of Opportunity

Continuing with Section 2.5.1, we demonstrate how to apply our method to enforce ε-equality of opportunity. To keep things simple, we assume there are only two demographic groups; *i.e.* $|\mathcal{A}| = 2$. Without loss of generality, we refer to one group as advantaged ($A = 1$) and the other as disadvantaged ($A = 0$). First we estimate rates $\mathbb{P}\{A = 1, Y = 1\}$ and $\mathbb{P}\{A = 0, Y = 1\}$ consistently by

$$\widehat{p}_1 = \tfrac{1}{n}\sum_{i=1}^n \mathbf{1}\{A_i = 1, Y_i = 1\} \quad \text{and} \quad \widehat{p}_0 = \tfrac{1}{n}\sum_{i=1}^n \mathbf{1}\{A_i = 0, Y_i = 1\}.$$

Then, we construct a robust constraint

$$\sup_{P: D_\varphi(P\|\widehat{P}_n)\le\rho/n} \mathbb{E}_{(X,A,Y)\sim P}\left[\tfrac{\mathbf{1}\{\widehat{Y}=1,A=1,Y=1\}}{\widehat{p}_1} - \tfrac{\mathbf{1}\{\widehat{Y}=1,A=0,Y=1\}}{\widehat{p}_0} - \varepsilon\right] \le 0,$$

or equivalently

$$\sup_{\boldsymbol{p}:\sum_{i=1}^n \varphi(np_i)\le\rho} \sum_{i=1}^n p_i\left[\tfrac{\mathbf{1}\{f_\theta(X_i)=1,A_i=1,Y_i=1\}}{\widehat{p}_1} - \tfrac{\mathbf{1}\{f_\theta(X_i)=1,A_i=0,Y_i=1\}}{\widehat{p}_0} - \varepsilon\right] \le 0.$$

A.7 Two-stage Method for Unknown Active Set

In this section, we show that the two-stage method in Section 2.5.1 also has the calibration property (similar to Theorem 2.3.1 and Corollary 2.3.1) if the true program $(\mathcal{P}_0)$, *i.e.* (2.3.1), is not ill-behaved.

First we recall the two-stage method:

1. At the first stage, we solve the program $(\mathcal{P}_n)$ with $\boldsymbol{\rho} = \mathbf{0}_K$, that is, the sample average approximation (SAA) problem. We identify that $\widehat{J}_{+,n} \subset [K]$ is the active set of the SAA problem, that is, the j-th constraint of the SAA problem is active if and only if $j \in \widehat{J}_{+,n}$.

2. At the second stage, we solve the program $(\mathcal{P}_n)$ with $\boldsymbol{\rho}$ such that ρ_j is some positive number if $j \in \widehat{J}_{+,n}$. This means that at the second stage we replace the sample mean approximation to the constraint by its distributionally robust counterpart if such constraint was identified as active at the first stage.

The index set $[K]$ can be partitioned into three parts:

$$J_+(\theta^\star, \boldsymbol{\lambda}^\star) = \{k \in [K] : \mathbb{E}_{P_0}[g_k(\theta^\star; Z)] = 0, \lambda_k^\star > 0\},$$
$$J_0(\theta^\star, \boldsymbol{\lambda}^\star) = \{k \in [K] : \mathbb{E}_{P_0}[g_k(\theta^\star; Z)] = 0, \lambda_k^\star = 0\},$$
$$J_-(\theta^\star, \boldsymbol{\lambda}^\star) = \{k \in [K] : \mathbb{E}_{P_0}[g_k(\theta^\star; Z)] < 0, \lambda_k^\star = 0\},$$

where J_+ is the active set with positive Lagrange multipliers, J_0 is the active set with zero Lagrange multipliers, and J_- is the inactive set. We assume the *strict complementarity* holds in the sense that $J_0(\theta^\star, \boldsymbol{\lambda}^\star) = \varnothing$.

Proposition A.7.1 (Preservance of active constraints)**.** *Assume the strict complementarity holds. We have* $\widehat{J}_{+,n} = J_+(\theta^\star, \boldsymbol{\lambda}^\star)$ *with probability converging to* 1 *as* $n \to \infty$.

Proof of Proposition A.7.1. Let $\widehat{\theta}_n^{(\mathrm{SAA})}$ be a solution to the sample average approximation problem at the first stage. Let $\widehat{\boldsymbol{\lambda}}_n$ be the associated Lagrange multiplier. It is known by the consistency of SAA Rubinstein and Shapiro (1993) that

$$\widehat{\lambda}_{k,n} = \lambda_k^\star + o_p(1) \text{ for } k \in J_+(\theta^\star, \boldsymbol{\lambda}^\star)$$

and

$$\mathbb{E}_P[g_k(\widehat{\theta}_n^{(\mathrm{SAA})}; Z)] = \mathbb{E}_P[g_k(\theta^\star; Z)] + o_p(1) \text{ for } k \in J_-(\theta^\star, \boldsymbol{\lambda}^\star)$$

Therefore, with probability converging to 1, $\widehat{\lambda}_{k,n} > 0$, $k \in J_+(\theta^\star, \boldsymbol{\lambda}^\star)$ and $\mathbb{E}_P[g_k(\widehat{\theta}_n^{(\mathrm{SAA})}; Z)] < 0$, $k \in J_-(\theta^\star, \boldsymbol{\lambda}^\star)$. Hence we complete the proof of Proposition A.7.1. □

Theorem A.7.1. *Suppose the true program $(\mathcal{P}_0)$ has m active constraints with positive Lagrange multipliers and $K-m$ inactive constraints (without loss of generality we let $J_+ = \{1,\ldots,m\}$ and $J_- = \{m+1,\ldots,K\}$). Let $\widehat{\theta}_n$ be the two-stage estimator. Under the standing assumptions, we have*

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}\left\{ \begin{bmatrix} \mathbb{E}_{P_0}[g_1(\widehat{\theta}_n; Z)] \\ \vdots \\ \mathbb{E}_{P_0}[g_K(\widehat{\theta}_n; Z)] \end{bmatrix} \in -\mathbb{R}_+^K \right\} \\
&= \mathbb{P}\left\{ \mathcal{N}\left(\mathbf{0}_m, \mathrm{Corr}_{P_0} \begin{bmatrix} g_1(\theta^\star; Z) \\ \vdots \\ g_m(\theta^\star; Z) \end{bmatrix} \right) \leq \begin{bmatrix} \sqrt{\rho_1} \\ \vdots \\ \sqrt{\rho_m} \end{bmatrix} \right\}.
\end{aligned}$$

This result shows that the limiting probability of satisfying the true constraints only depends on the correlation structure between active constraints and the uncertainty set radii for the active constraints.

Proof of Theorem A.7.1. At the first stage, we identify $J \subset [K]$ as active set with probability p_J. Here the randomness is introduced by the data samples $Z_1,\ldots,Z_n$. Let $2^{[K]}$ be the power set of $[K]$. We have

$$\sum_{J\in 2^{[K]}} p_J = 1.$$

By Proposition A.7.1, as $n \to \infty$ we have

$$p_{J_+} \to 1$$

and

$$p_J \to 0 \text{ for any } J \in 2^{[K]} \text{ and } J \neq J_+.$$

By Corollary 2.3.1, we have

$$\lim_{n\to\infty} \mathbb{P}\left\{\widehat{\theta}_n \text{ is feasible} \mid \text{identify } J_+ \text{ as active set}\right\}$$
$$= \mathbb{P}\left\{\mathcal{N}\left(\mathbf{0}_m, \mathrm{Corr}_{P_0}\begin{bmatrix} g_1(\theta^\star; Z) \\ \vdots \\ g_m(\theta^\star; Z)\end{bmatrix}\right) \le \begin{bmatrix}\sqrt{\rho_1} \\ \vdots \\ \sqrt{\rho_m}\end{bmatrix}\right\}.$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left\{\widehat{\theta}_n \text{ is feasible}\right\} \\
&= \sum_{J\in 2^{[K]}} \mathbb{P}\left\{\widehat{\theta}_n \text{ is feasible} \mid \text{identify } J \text{ as active set}\right\} \mathbb{P}\left\{\text{identify } J \text{ as active set}\right\} \\
&= \underbrace{p_{J_+}}_{\to 1} \mathbb{P}\left\{\widehat{\theta}_n \text{ is feasible} \mid \text{identify } J_+ \text{ as active set}\right\} \\
&\quad + \sum_{J\neq J_+} \underbrace{p_J \mathbb{P}\left\{\widehat{\theta}_n \text{ is feasible} \mid \text{identify } J \text{ as active set}\right\}}_{\to 0} \\
&\to \mathbb{P}\left\{\mathcal{N}\left(\mathbf{0}_m, \mathrm{Corr}_{P_0}\begin{bmatrix} g_1(\theta^\star; Z) \\ \vdots \\ g_m(\theta^\star; Z)\end{bmatrix}\right) \le \begin{bmatrix}\sqrt{\rho_1} \\ \vdots \\ \sqrt{\rho_m}\end{bmatrix}\right\}.
\end{aligned}$$

Hence we complete the proof of Theorem A.7.1. □

A.8 Proxy Dual Ascent Algorithm for Non-differentiable Constraints

We summarize in Algorithm 5 the proxy dual ascent algorithm for solving a stochastic optimization problem with single non-differentiable constraint. The difference between Algorithm 5 and 1 is in the step of evaluating (proxy) dual function: Algorithm 5 uses a differentiable surrogate $\tilde{g}$ (highlighted in orange) instead of the

non-differentiable g in Algorithm 1 in this step.

Algorithm 5 Proxy dual ascent algorithm for single non-differentiable constraint

1: **Input:** starting dual iterate $\lambda_0 \geq 0$
2: **repeat**
3: Evaluate proxy dual function:

$$(\theta_t, \mu_t, \nu_t) \leftarrow \arg\min_{\theta, \mu \geq 0, \nu} \frac{1}{n} \sum_{i=1}^n f(\theta; Z_i) + \lambda_t \left\{ \frac{1}{n} \sum_{i=1}^n \mu \varphi^* \left(\frac{\tilde{g}(\theta; Z_i) - \nu}{\mu} \right) + \mu \frac{\rho_\alpha}{n} + \nu \right\}$$

4: Dual ascent update: $\lambda_{t+1} \leftarrow \left[\lambda_t + \eta_t \left\{ \frac{1}{n} \sum_{i=1}^n \mu_t \varphi^* \left(\frac{g(\theta_t; Z_i) - \nu_t}{\mu_t} \right) + \mu_t \frac{\rho_\alpha}{n} + \nu_t \right\} \right]_+$
5: **until** converged

Algorithm 6 summarizes the proxy dual ascent algorithm for solving a stochastic optimization problem with multiple non-differentiable constraints. In contrast to Algorithm 4, Algorithm 6 replaces the non-differentiable functions g_k's by their differentiable surrogates $\tilde{g}_k$'s (highlighted in orange) in the (proxy) dual function evaluation step.

Algorithm 6 Proxy dual ascent algorithm for multiple non-differentiable constraints

1: **Input:** starting dual iterate $\boldsymbol{\lambda}_0 = (\lambda_{01}, \ldots, \lambda_{0K})^\top \in \mathbb{R}_+^K$
2: **repeat**
3: Evaluate proxy dual function:

$$(\theta_t, \boldsymbol{\mu}_t, \boldsymbol{\nu}_t) \leftarrow \arg\min_{\theta, \boldsymbol{\mu} \in \mathbb{R}_+^K, \boldsymbol{\nu}} \Big(\frac{1}{n} \sum_{i=1}^n f(\theta; Z_i)$$

$$+ \sum_{k=1}^K \lambda_{tk} \left\{ \frac{1}{n} \sum_{i=1}^n \mu_{tk} \varphi^* \left(\frac{\tilde{g}_k(\theta; Z_i) - \nu_{tk}}{\mu_{tk}} \right) + \mu_{tk} \frac{\rho_k}{n} + \nu_{tk} \right\} \Big)$$

4: Dual ascent update:

$$\lambda_{t+1,k} \leftarrow \left[\lambda_{tk} + \eta_t \left\{ \frac{1}{n} \sum_{i=1}^n \mu_{tk} \varphi^* \left(\frac{g_k(\theta_t; Z_i) - \nu_{tk}}{\mu_{tk}} \right) + \mu_{tk} \frac{\rho_\alpha}{n} + \nu_{tk} \right\} \right]_+, k \in [K]$$

5: **until** converged

A.9 Real Data Experiments: Details and More

In this section, we provide details for experiments on Adult dataset with ε-demographic parity as the fairness goal (which we present in the main text), and

conduct additional experiments on Adult dataset with ε-false positive rate parity as the fairness goal.

A.9.1 Adult with ε-Demographic Parity

Recall that in Section 2.5.3 the fairness goal is ε-demographic parity (ε-DP):

$$\left|\mathbb{P}(\widehat{Y} = 1 \mid A = 1) - \mathbb{P}(\widehat{Y} = 1 \mid A = 0)\right| \leq \varepsilon,$$

where $A = 1$ for male is the advantaged group and $A = 0$ for female is the disadvantaged group. We use a logistic regression model for classification by predicting $\widehat{Y} = \mathbf{1}\{\theta^\top X \geq 0\}$ and training such model parameterized in θ by logistic loss. We implement the two-stage method and proxy dual ascent algorithm by replacing indicator function by the sigmoidal function $h_1(t) = (1 + e^{-at})^{-1}$ for $a = 2$. We do bootstrap evaluation by treating the full dataset as the true probability measure and sample such probability measure with replacement as the training distribution (with sampling rate 50%).

A.9.2 Adult with ε-False Positive Rate Parity

Similar to the preceding subsection, now we consider the fairness goal to be ε-false positive rate parity:

$$\left|\mathbb{P}(\widehat{Y} = 1 \mid Y = 0, A = 1) - \mathbb{P}(\widehat{Y} = 1 \mid Y = 0, A = 0)\right| \leq \varepsilon,$$

where $A = 1$ for male is the advantaged group and $A = 0$ for female is the disadvantaged group.

In Figure A.3, we have line plots for frequency of constraint satisfaction and box plots for classification error rate, all of which are summarized over 100 replicates. The patterns shown in the left and right panels are similar to that in Figure II.2. We

observe more variation in frequencies of constraint satisfaction across different false positive rate tolerance designs due to the label and demographic attribute imbalance of Adult dataset.

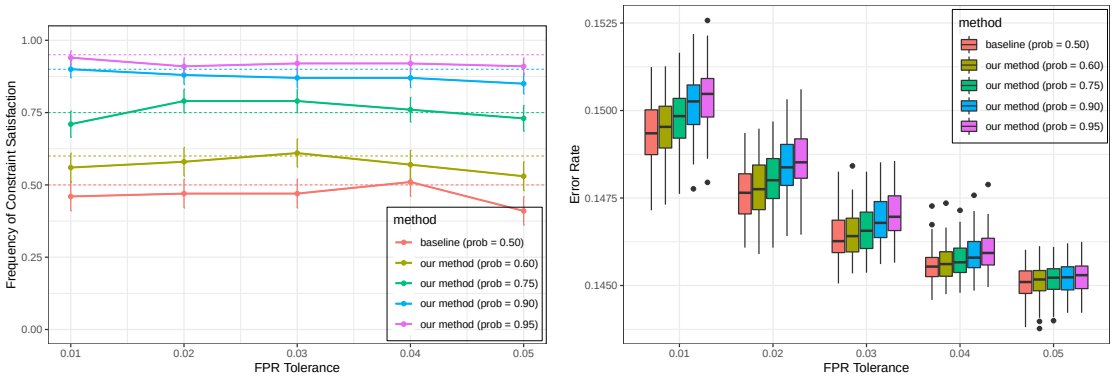

Figure A.3: Frequency of constraint satisfaction (left) and classification error rate (right) for different false positive rate parity tolerance $\varepsilon \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$. Baseline (sample average approximation, SAA) and our methods (with nominal probability $0.60, 0.75, 0.90, 0.95$) are compared.

A.10 Dual Form of the Robust Constraint Function (2.2.5)

In this section, we provide a standard derivation for the dual form of the robust constraint function (2.2.5).

We introduce the likelihood ratio $L(Z) = dP(Z)/d\widehat{P}_n(Z)$. By change of variable, we can rewrite the robust constraint function (2.2.5) as

$$\begin{aligned} & \sup_{P: D_\varphi(P\|\widehat{P}_n)\le\rho} \mathbb{E}_P\big[g(\theta;Z)\big] \\ = & \sup_{L\ge 0} \left\{ \mathbb{E}_{\widehat{P}_n}\big[L(Z)g(\theta;Z)\big] \mid \mathbb{E}_{\widehat{P}_n}\big[\varphi(L(Z))\big] \le \rho, \mathbb{E}_{\widehat{P}_n}\big[L(Z)\big] = 1 \right\}, \end{aligned}$$

where the supremum takes over measurable functions. This gives us a constrained optimization problem. Let $\mu \ge 0$ be the Lagrange multiplier for $\mathbb{E}_{\widehat{P}_n}\big[\varphi(L(Z))\big] \le \rho$ and $\nu \in \mathbb{R}$ be the Lagrange multiplier for $\mathbb{E}_{\widehat{P}_n}\big[L(Z)\big] = 1$. The corresponding

Lagrangian is

$$\mathcal{L}(L,\mu,\nu)=\mathbb{E}_{\widehat{P}_n}\left[(g(\theta;Z)-\nu)L(Z)-\mu\varphi(L(Z))\right]+\mu\rho+\nu.$$

For regular φ-divergence, we have

$$\begin{aligned}
&\sup_{P:D_\varphi(P\|\widehat{P}_n)\le\rho}\mathbb{E}_P\left[g(\theta;Z)\right]\\
&=\inf_{\mu\ge0,\nu\in\mathbb{R}}\sup_{L\ge0}\mathcal{L}(L,\mu,\nu)\\
&=\inf_{\mu\ge0,\nu\in\mathbb{R}}\sup_{L\ge0}\left\{\sum_{i=1}^n\left[(g(\theta;Z_i)-\nu)L(Z_i)-\mu\varphi(L(Z_i))\right]+\mu\rho+\nu\right\}\\
&=\inf_{\mu\ge0,\nu\in\mathbb{R}}\sup_{L\ge0}\left\{\sum_{i=1}^n\mu\left[\frac{g(\theta;Z_i)-\nu}{\mu}L(Z_i)-\varphi(L(Z_i))\right]+\mu\rho+\nu\right\}\\
&=\inf_{\mu\ge0,\nu\in\mathbb{R}}\left\{\sum_{i=1}^n\mu\sup_{t_i\ge0}\left\{\frac{g(\theta;Z_i)-\nu}{\mu}t_i-\varphi(t_i)\right\}+\mu\rho+\nu\right\}\\
&=\inf_{\mu\ge0,\nu\in\mathbb{R}}\left\{\sum_{i=1}^n\mu\varphi^*\left(\frac{g(\theta;Z_i)-\nu}{\mu}\right)+\mu\rho+\nu\right\}.
\end{aligned}$$

Here the last equality holds according to the definition of the convex conjugate $\varphi^*(\cdot)$.

A.11 Experiments on Additional Baseline and Dataset

In this section, we conduct more experiments using two-dataset approach of (Cotter et al., 2019a) as an additional baseline and default of credit card clients dataset from UCI (Dua and Graff, 2017) as an additional dataset.

A.11.1 Adult with ε-Demographic Parity (Continued)

We continue the experiments on Adult dataset with ε-demographic parity in Section 2.5.3 and A.9.1 by adding two-dataset approach of (Cotter et al., 2019a) as an additional baseline.

The two dataset approach of (Cotter et al., 2019a) splits the training set into

two parts, one for updating the model parameters and the other for updating the Lagrangian multipliers, with the goal to improve the generalization of fairness constraints. Figure A.4 demonstrates that the two-dataset approach marginally improves the frequency of constraint satisfaction over the SAA. The cost of such an improvement is increased variation in classification error rate. Due to the fact that the two-dataset approach does not use the entire training set to update the model parameters, statistical efficiency is sacrificed. Although the two-dataset approach outperforms the SAA in terms of constraint satisfaction frequency, both are inferior to our methods, which achieve the user-prescribed level of frequency.

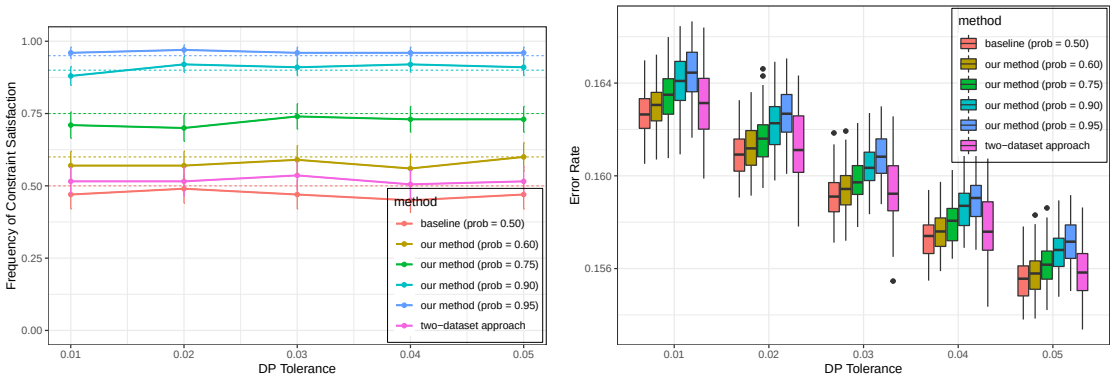

Figure A.4: Frequency of constraint satisfaction (left) and classification error rate (right) for different demographic parity tolerance $\varepsilon \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$. Baseline (sample average approximation, SAA), our methods (with nominal probability $0.60, 0.75, 0.90, 0.95$), and two-dataset approach (Cotter et al., 2019a) are compared.

A.11.2 Adult with ε-False Positive Rate Parity (Continued)

We continue the experiments on Adult dataset with ε-false positive rate parity in Section A.9.2 by adding two-dataset approach of (Cotter et al., 2019a) as an additional baseline.

Figure A.5 demonstrates similar patterns as Figure A.4. The two-dataset approach improves constraint satisfaction frequency over the SAA, but has a worse classification error rate or one comparable to it. Our methods achieve desired frequency of

constraint satisfaction at the user-prescribed level.

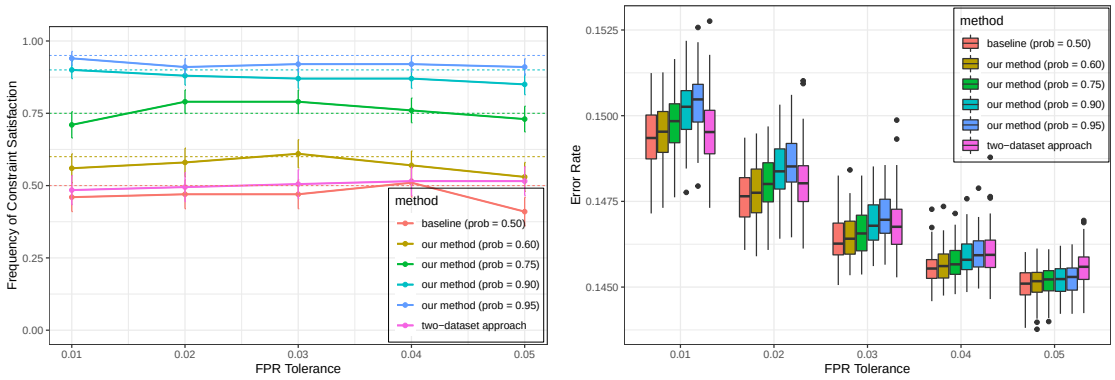

Figure A.5: Frequency of constraint satisfaction (left) and classification error rate (right) for different false positive rate parity tolerance $\varepsilon \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$. Baseline (sample average approximation, SAA), our methods (with nominal probability $0.60, 0.75, 0.90, 0.95$), and two-dataset approach (Cotter et al., 2019a) are compared.

A.11.3 Credit with ε-Demographic Parity

Predicting whether or not an individual defaulted on a loan is the classification task of the UCI default of credit card clients (Credit) dataset. Membership in the demographic group is determined by an individual's level of education: $A = 1$ if a person has earned a graduate degree; otherwise, $A = 0$.

We consider the fairness goal to be ε-demographic parity:

$$\left|\mathbb{P}(\widehat{Y} = 1 \mid A = 1) - \mathbb{P}(\widehat{Y} = 1 \mid A = 0)\right| \leq \varepsilon,$$

where $A = 1$ for individuals with a graduate degree is the advantaged group.

As depicted in Figure A.6, our methods achieve the level of constraint satisfaction frequency specified by the user at the expense of a slight increase in classification error.

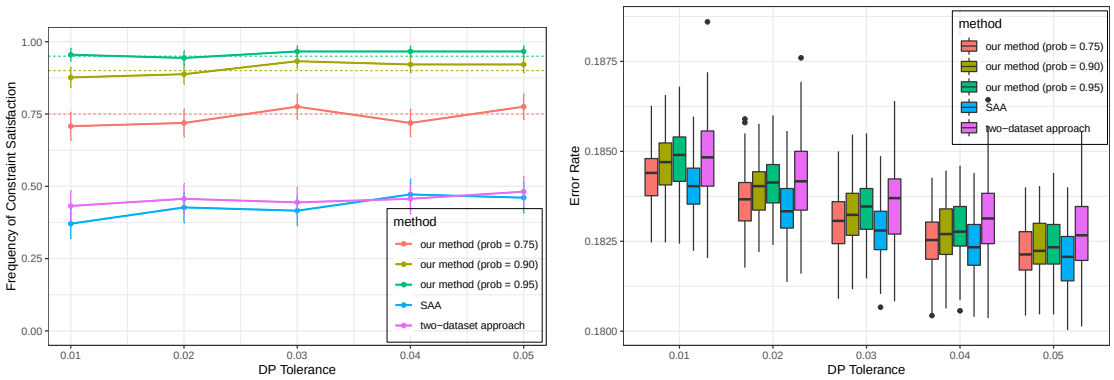

Figure A.6: Frequency of constraint satisfaction (left) and classification error rate (right) for different demographic parity tolerance $\varepsilon \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$. Our methods (with nominal probability 0.75, 0.90, 0.95), sample average approximation (SAA) and two-dataset approach (Cotter et al., 2019a) are compared.

A.11.4 Credit with ε-True Positive Rate Parity

Similar to the preceding subsection, now we consider the fairness goal to be ε-true positive rate parity:

$$\left|\mathbb{P}(\widehat{Y} = 1 \mid Y = 1, A = 1) - \mathbb{P}(\widehat{Y} = 1 \mid Y = 1, A = 0)\right| \leq \varepsilon,$$

where $A = 1$ for individuals with a graduate degree is the advantaged group.

The patterns shown in Figure A.7 are similar to that in Figure A.6. The SAA has the lowest error rate but the worst generalization of constraint satisfaction. The two-dataset approach increases the SAA's frequency of constraint satisfaction at the cost of an increase in classification error rates. Our methods can achieve the desired high probability of constraint satisfaction, whereas neither of the two baselines can.

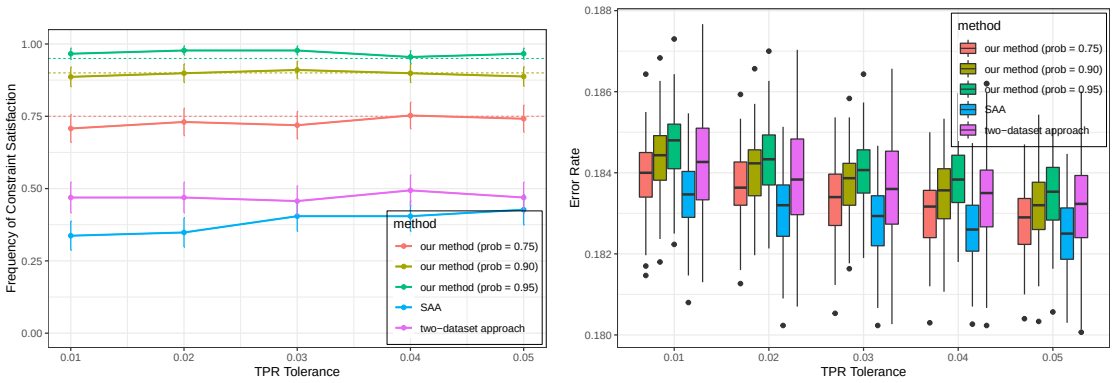

Figure A.7: Frequency of constraint satisfaction (left) and classification error rate (right) for different true positive rate parity tolerance $\varepsilon \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$. Our methods (with nominal probability 0.75, 0.90,0.95), sample average approximation (SAA) and two-dataset approach (Cotter et al., 2019a) are compared.

APPENDIX B

Appendix of Chapter III

In Section B.1 and Section B.2, we have more discussions on the privacy cost of linear fairness and linear-fractional fairness, respectively. In Section B.3, we present the proof of Theorem 3.3.2. In Section B.4 and Section B.5, we provide additional treatments to multiple demographic groups and missing sensitive attributes.

B.1 Linear Fairness Constraint

We extend the notion of demographic parity to a more general form: we say that θ is fair (with respect to A) if

$$\mathbb{E}\big[g(\theta; X, Y)|A = 1\big] - \mathbb{E}\big[g(\theta; X, Y)|A = 0\big] = 0. \tag{B.1.1}$$

The fairness notion (B.1.1) is known as *linear* fairness constraint (Celis et al., 2021). Note that demographic parity is a special case of (B.1.1) if we take $g(\theta; X, Y) = \mathbf{1}\{f_\theta(X) = 1\}$.

On the one hand, enforcing fairness notion (B.1.1) with respect to A is

$$\mathbb{E}_{(X,Y)|A=1}\big[g(\theta; X, Y)\big] - \mathbb{E}_{(X,Y)|A=0}\big[g(\theta; X, Y)\big] = 0$$

or equivalently

$$\mathbb{E}_{Q_1}\big[g(\theta; X, Y)\big] - \mathbb{E}_{Q_0}\big[g(\theta; X, Y)\big] = 0.$$

On the other hand, enforcing fairness notion (B.1.1) with respect to Z is

$$\mathbb{E}_{(X,Y)|Z=1}\big[g(\theta; X, Y)\big] - \mathbb{E}_{(X,Y)|Z=0}\big[g(\theta; X, Y)\big] = 0$$

or equivalently

$$\mathbb{E}_{\frac{\gamma\pi_0}{\gamma\pi_0+(1-\gamma)\pi_1}Q_0+\frac{(1-\gamma)\pi_1}{\gamma\pi_0+(1-\gamma)\pi_1}Q_1}\big[g(\theta; X, Y)\big] - \mathbb{E}_{\frac{(1-\gamma)\pi_0}{(1-\gamma)\pi_0+\gamma\pi_1}Q_0+\frac{\gamma\pi_1}{(1-\gamma)\pi_0+\gamma\pi_1}Q_1}\big[g(\theta; X, Y)\big] = 0.$$

Therefore, the true fairness constraint function is

$$c(\theta) = \int_{\mathcal{X}\times\mathcal{Y}} g(\theta; x, y) d(Q_1 - Q_0)(x, y),$$

while the proxy fairness constraint function is

$$\begin{aligned}
\widetilde{c}(\theta) &= \left(-\frac{\gamma\pi_0}{\gamma\pi_0 + (1-\gamma)\pi_1} + \frac{(1-\gamma)\pi_0}{(1-\gamma)\pi_0 + \gamma\pi_1}\right) \int_{\mathcal{X}\times\mathcal{Y}} g(\theta; x, y) d(Q_1 - Q_0)(x, y) \\
&= \left(\frac{(1-\gamma)\pi_1}{\gamma\pi_0 + (1-\gamma)\pi_1} - \frac{\gamma\pi_1}{(1-\gamma)\pi_0 + \gamma\pi_1}\right) \int_{\mathcal{X}\times\mathcal{Y}} g(\theta; x, y) d(Q_1 - Q_0)(x, y) \\
&\triangleq \psi_{\text{lin}}(\gamma, \pi_0, \pi_1) \times c(\theta).
\end{aligned} \tag{B.1.2}$$

By (B.1.2), the proxy constraint function $\widetilde{c}(\theta)$ is equal to the true $c(\theta)$ up to a scaling factor

$$\begin{aligned}
\psi_{\text{lin}}(\gamma, \pi_0, \pi_1) &\triangleq -\frac{\gamma\pi_0}{\gamma\pi_0 + (1-\gamma)\pi_1} + \frac{(1-\gamma)\pi_0}{(1-\gamma)\pi_0 + \gamma\pi_1} \\
&= \frac{(1-\gamma)\pi_1}{\gamma\pi_0 + (1-\gamma)\pi_1} - \frac{\gamma\pi_1}{(1-\gamma)\pi_0 + \gamma\pi_1} \\
&= \frac{(1-2\gamma)\pi_0\pi_1}{\{\gamma\pi_0 + (1-\gamma)\pi_1\}\{(1-\gamma)\pi_0 + \gamma\pi_1\}}.
\end{aligned} \tag{B.1.3}$$

This also implies $c(\theta) = 0$ if and only if $\widetilde{c}(\theta) = 0$, providing an alternative proof for

Proposition 3.3.1.

Now we are ready to show the privacy cost in linear fairness (B.1.1)-aware learning. First, let the true parameter $\theta^\star$, *i.e.* the solution to the population problem, be

$$\theta^\star \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \mathbb{E}\big[\ell(\theta; X, Y)\big] \\ \text{subject to} & \mathbb{E}\big[g(\theta; X, Y)|A=1\big] - \mathbb{E}\big[g(\theta; X, Y)|A=0\big] = 0 \end{array} \right\}, \tag{B.1.4}$$

where the expectation is with respect to the underlying distribution of tuple (X, A, Y).

Then, let the estimator $\widehat{\theta}_n$ be the solution to the empirical problem given the true sensitive attribute,

$$\widehat{\theta}_n \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \frac{1}{n}\sum_{i=1}^n \ell(\theta; X_i, Y_i) \\ \text{subject to} & \left| \frac{\sum_{i=1}^n g(\theta; X_i, Y_i)\mathbf{1}\{A_i=1\}}{\sum_{i=1}^n \mathbf{1}\{A_i=1\}} - \frac{\sum_{i=1}^n g(\theta; X_i, Y_i)\mathbf{1}\{A_i=0\}}{\sum_{i=1}^n \mathbf{1}\{A_i=0\}} \right| \le \alpha_n \end{array} \right\}.$$

Finally, let the estimator $\widetilde{\theta}_n$ be the solution to the empirical problem given the proxy sensitive attribute,

$$\widetilde{\theta}_n \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \frac{1}{n}\sum_{i=1}^n \ell(\theta; X_i, Y_i) \\ \text{subject to} & \left| \frac{\sum_{i=1}^n g(\theta; X_i, Y_i)\mathbf{1}\{Z_i=1\}}{\sum_{i=1}^n \mathbf{1}\{Z_i=1\}} - \frac{\sum_{i=1}^n g(\theta; X_i, Y_i)\mathbf{1}\{Z_i=0\}}{\sum_{i=1}^n \mathbf{1}\{Z_i=0\}} \right| \le \alpha_n \end{array} \right\}.$$

We made the following technical assumptions on the problem (B.1.4).

1. **smoothness and concentration:** ℓ and g are twice continuously differentiable with respect to θ, and $\ell(\theta^\star; X, Y)$, $\nabla\ell(\theta^\star; X, Y)$, $g(\theta^\star; X, Y)$, $\nabla g(\theta^\star; X, Y)$ are sub-Gaussian random variables.

2. **uniqueness:** the stochastic optimization problem with a single expected value constraint (B.1.4) has a unique optimal primal-dual pair $(\theta^\star, \lambda^\star)$, and $\theta^\star$ belongs to the interior of the compact set Θ.

3. **positive definiteness:** The Hessian of the Lagrangian evaluated at $(\theta^\star, \lambda^\star)$ is

positive definite.

We have the main technical result as follows.

Theorem B.1.1 (Privacy cost in linear fairness (B.1.1)-aware learning)**.** *Under the standing assumptions, let estimators $\widehat{\theta}_n$ and $\widetilde{\theta}_n$ be consistent for $\theta^\star$, then*

$$\sqrt{n}\{c(\widehat{\theta}_n) - \cancel{c(\theta^\star)}\} \xrightarrow{d} \mathcal{N}(0, \sigma^2) \text{ and } \sqrt{n}\{c(\widetilde{\theta}_n) - \cancel{c(\theta^\star)}\} \xrightarrow{d} \mathcal{N}(0, \widetilde{\sigma}^2),$$

where

$$\sigma^2 = \frac{\mathrm{Var}_{Q_0}[g(\theta^\star; X, Y)]}{\pi_0} + \frac{\mathrm{Var}_{Q_1}[g(\theta^\star; X, Y)]}{\pi_1}$$

and

$$\widetilde{\sigma}^2 = \psi_{\mathrm{lin}}^{-2} \times \left\{ \frac{\mathrm{Var}_{\widetilde{Q}_0}[g(\theta^\star; X, Y)]}{\widetilde{\pi}_0} + \frac{\mathrm{Var}_{\widetilde{Q}_1}[g(\theta^\star; X, Y)]}{\widetilde{\pi}_1} \right\}.$$

The asymptotic relative efficiency (ARE) of $\widetilde{\theta}_n$ to $\widehat{\theta}_n$ is

$$\mathrm{ARE}(\widetilde{\theta}_n, \widehat{\theta}_n) = \varphi\left(\gamma, \frac{\pi_0}{\pi_1}, \frac{\mathrm{Var}[g(\theta^\star; X, Y)|A = 0]}{\mathrm{Var}[g(\theta^\star; X, Y)|A = 1]}\right),$$

where

$$\varphi(\gamma, r_1, r_2) \triangleq \frac{(1-2\gamma)^2 r_1(r_1 + r_2)}{\{\gamma r_1 + (1-\gamma)\}^2\{(1-\gamma)r_1 r_2 + \gamma\} + \{(1-\gamma)r_1 + \gamma\}^2\{\gamma r_1 r_2 + (1-\gamma)\}}.$$

Proof of Theorem B.1.1. Note that Theorem 3.3.2 implies Theorem B.1.1 by letting $h(X, Y) \equiv 1$. Therefore, it is sufficient to prove Theorem 3.3.2, whose proof can be found in Appendix B.3. □

Theorem B.1.1 demonstrates that the cost of privacy is the efficiency loss in terms of fairness violations. For fixed ratios $r_1 \triangleq \pi_0/\pi_1 > 0$ and $r_2 \triangleq \mathrm{Var}[g(\theta^\star; X, Y)|A = 0]/\mathrm{Var}[g(\theta^\star; X, Y)|A = 1] > 0$, $\varphi_{\mathrm{lin}}(\gamma, r_1, r_2)$ is a decreasing function in γ. In the absence of privacy, $\varphi_{\mathrm{lin}}(0, r_1, r_2) = 1$ means no efficiency loss. Under perfect privacy, $\varphi_{\mathrm{lin}}(0.5, r_1, r_2) = 0$ indicates total loss of efficiency. Moreover, $\widehat{\theta}_n$ is always more

efficient than $\widetilde{\theta}_n$ because $\mathrm{ARE}(\widetilde{\theta}_n, \widehat{\theta}_n) \leq 1$.

Figure B.1 demonstrates the asymptotic relative efficiency (ARE) curve of privacy level γ for varying ratios r_1 and r_2. The ARE is always upper bounded by $(1-2\gamma)^2$, which is achieved only if $\pi_0 = \pi_1 = 0.5$. Recall that $\pi_0 = \mathbb{P}(A=0)$ and $\pi_1 = \mathbb{P}(A=1)$. Therefore for any fixed γ and r_2, the ARE achieves its maximum only if the dataset is balanced in the sensitive attribute A. Moreover, for any fixed γ and r_2, the ARE is strictly increasing in π_0 (assuming $\pi_0 < 0.5$). This implies the effect of subgroup size imbalance: demographic group imbalance degrades the efficiency loss in privately fair learning. Lastly, the ARE is strictly increasing in the problem-specific parameter r_2, given fixed γ and $r_1 < 1$.

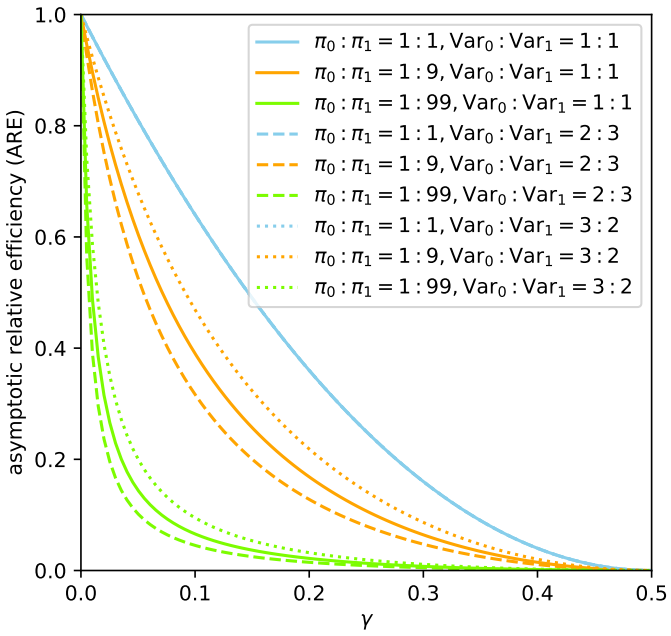

Figure B.1: Asymptotic relative efficiency curve of γ for varying ratios of π_0 to π_1 and $\mathrm{Var}[g(\theta^\star; X, Y)|A=0]$ to $\mathrm{Var}[g(\theta^\star; X, Y)|A=1]$.

B.2 Linear-fractional Fairness Constraint

We provide further discussion to supplement Section 3.3. Recall the marginal distributions and conditional distributions in (3.3.2) and

$$\begin{cases} \mathbb{P}(Z=0)=\widetilde{\pi}_0, & \mathbb{P}(Z=1)=\widetilde{\pi}_1 \\ (X,Y)|Z=0\sim\widetilde{Q}_0, & (X,Y)|Z=1\sim\widetilde{Q}_1 \end{cases}.$$

Under the private mechanism Q in (3.3.6), we have

$$\begin{cases} \widetilde{\pi}_0=(1-\gamma)\pi_0+\gamma\pi_1, \widetilde{\pi}_1=\gamma\pi_0+(1-\gamma)\pi_1 \\ \widetilde{Q}_0 \overset{d}{=} \frac{(1-\gamma)\pi_0}{(1-\gamma)\pi_0+\gamma\pi_1}Q_0+\frac{\gamma\pi_1}{(1-\gamma)\pi_0+\gamma\pi_1}Q_1 \\ \widetilde{Q}_1 \overset{d}{=} \frac{\gamma\pi_0}{\gamma\pi_0+(1-\gamma)\pi_1}Q_0+\frac{(1-\gamma)\pi_1}{\gamma\pi_0+(1-\gamma)\pi_1}Q_1 \end{cases}. \tag{B.2.1}$$

The marginal distribution and conditional distribution in (B.2.1) uniquely identify the joint distribution of (X,Z,Y).

On the one hand, enforcing fairness notion (3.3.1) with respect to A is

$$\frac{\mathbb{E}_{(X,Y)|A=1}\big[g(\theta;X,Y)\big]}{\mathbb{E}_{(X,Y)|A=1}\big[h(X,Y)\big]}-\frac{\mathbb{E}_{(X,Y)|A=0}\big[g(\theta;X,Y)\big]}{\mathbb{E}_{(X,Y)|A=0}\big[h(X,Y)\big]}=0$$

or equivalently

$$c(\theta)\triangleq\frac{\mathbb{E}_{Q_1}\big[g(\theta;X,Y)\big]}{\mathbb{E}_{Q_1}\big[h(X,Y)\big]}-\frac{\mathbb{E}_{Q_0}\big[g(\theta;X,Y)\big]}{\mathbb{E}_{Q_0}\big[h(X,Y)\big]}=0.$$

On the other hand, enforcing fairness notion (3.3.1) with respect to Z is

$$\frac{\mathbb{E}_{(X,Y)|Z=1}\big[g(\theta;X,Y)\big]}{\mathbb{E}_{(X,Y)|Z=1}\big[h(X,Y)\big]}-\frac{\mathbb{E}_{(X,Y)|Z=0}\big[g(\theta;X,Y)\big]}{\mathbb{E}_{(X,Y)|Z=0}\big[h(X,Y)\big]}=0$$

or equivalently

$$\widetilde{c}(\theta) \triangleq \left\{ \begin{array}{c} \dfrac{\gamma\pi_0\mathbb{E}_{Q_0}\big[g(\theta;X,Y)\big] + (1-\gamma)\pi_1\mathbb{E}_{Q_1}\big[g(\theta;X,Y)\big]}{\gamma\pi_0\mathbb{E}_{Q_0}\big[h(X,Y)\big] + (1-\gamma)\pi_1\mathbb{E}_{Q_1}\big[h(X,Y)\big]} \\ -\dfrac{(1-\gamma)\pi_0\mathbb{E}_{Q_0}\big[g(\theta;X,Y)\big] + \gamma\pi_1\mathbb{E}_{Q_1}\big[g(\theta;X,Y)\big]}{(1-\gamma)\pi_0\mathbb{E}_{Q_0}\big[h(X,Y)\big] + \gamma\pi_1\mathbb{E}_{Q_1}\big[h(X,Y)\big]} \end{array} \right\} = 0.$$

By some algebra, we find that the proxy constraint function $\widetilde{c}(\theta)$ is equal to the true constraint function $c(\theta)$ up to a scaling factor: $\widetilde{c}(\theta) = \psi_{\text{frac}}(\gamma, \pi_0, \pi_1, m_0, m_1) \times c(\theta)$, where

$$\psi_{\text{frac}}(\gamma, \pi_0, \pi_1, m_0, m_1) \triangleq \frac{(1-2\gamma)\pi_0\pi_1 m_0 m_1}{\{\gamma\pi_0 m_0 + (1-\gamma)\pi_1 m_1\}\{(1-\gamma)\pi_0 m_0 + \gamma\pi_1 m_1\}}, \quad \text{(B.2.2)}$$

as well $m_0 \triangleq \mathbb{E}_{Q_0}\big[h(X,Y)\big]$ and $m_1 \triangleq \mathbb{E}_{Q_1}\big[h(X,Y)\big]$.

By comparing the scaling factor (B.2.2) with the functional form of (B.1.3), we can rewrite $\psi_{\text{frac}}(\cdot)$ by

$$\psi_{\text{frac}}(\gamma, \pi_0, \pi_1, m_0, m_1) = \psi_{\text{lin}}(\gamma, \pi_0 m_0, \pi_1 m_1).$$

Therefore, we can interpret the scaling factor $\psi_{\text{frac}}(\cdot)$ by treating $\pi_0 m_0$ and $\pi_1 m_1$ as a whole, allowing us to understand the privacy cost from a different perspective. Note that for equality of opportunity, we have $\pi_a m_a = \mathbb{P}(A=a)\mathbb{P}(Y=1|A=a) = \mathbb{P}(Y=1, A=a)$ for $a \in \{0,1\}$.

For equality of opportunity, Mozannar et al. (2020) show a sample complexity bound for the fairness violation of the estimator $\widetilde{\theta}_n$:

$$c(\widetilde{\theta}_n) - \cancel{c(\theta^\star)} \leq \frac{C_1(1-\gamma)}{(1-2\gamma)p^2}\left(C_2 + C_3\mathcal{R}_{\frac{np}{4}}(\mathcal{F}) + \frac{C_4}{\sqrt{n\delta p}}\right) \quad \text{(B.2.3)}$$

with probability at least $1-\delta$, where $p = \min\{\mathbb{P}(Y=1, A=0), \mathbb{P}(Y=1, A=1)\}$, $\mathcal{R}_\cdot(\cdot)$ is the Rademacher complexity, and C_i's $(1 \leq i \leq 4)$ are some universal

constants. Not precisely, the upper bound (B.2.3) reflects the effect of privacy level via γ and the effect of dataset imbalance through p. Comparing to this, our theory states that

$$\lim_{n\to\infty} \frac{\mathrm{Var}[c(\widehat{\theta}_n) - \cancel{c(\theta^\star)}]}{\mathrm{Var}[c(\widetilde{\theta}_n) - \cancel{c(\theta^\star)}]} = \varphi\left(\gamma, \frac{\mathbb{P}(Y=1, A=0)}{\mathbb{P}(Y=1, A=1)}, 1\right),$$

which is depicted by Figure B.2.

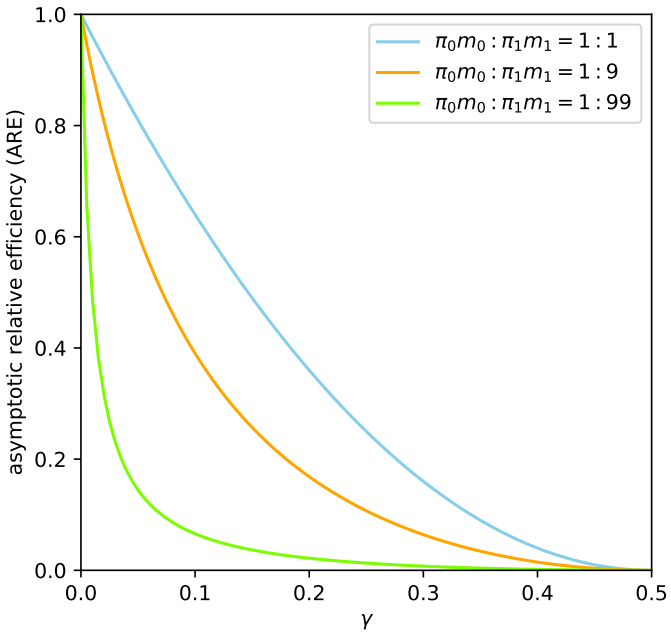

Figure B.2: Asymptotic relative efficiency curve of γ for varying ratio of $\mathbb{P}(Y=1, A=0)$ to $\mathbb{P}(Y=1, A=1)$.

B.3 Proof of Theorem 3.3.2

First, we prove the case when $\alpha_n = 0$ for all n. For this case both the population problem and the empirical problem are subject to equality constraints.

Consider a stochastic optimization problem with linear-fractional constraint

$$(\mathcal{P}_0): \quad \theta^\star \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \mathbb{E}\big[\ell(\theta; X, Y)\big] \\ \text{subject to} & \dfrac{\mathbb{E}\big[g(\theta; X, Y)|A=1\big]}{\mathbb{E}\big[h(X, Y)|A=1\big]} - \dfrac{\mathbb{E}\big[g(\theta; X, Y)|A=0\big]}{\mathbb{E}\big[h(X, Y)|A=0\big]} = 0 \end{array} \right\},$$

where the expectation is with respect to the underlying distribution of tuple (X, A, Y).

The corresponding empirical problem given the true sensitive attribute is

$$(\mathcal{P}_n): \quad \widehat{\theta}_n \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \frac{1}{n}\sum_{i=1}^n \ell(\theta; X_i, Y_i) \\ \text{subject to} & \frac{\sum_{i=1}^n g(\theta;X_i,Y_i)\mathbf{1}\{A_i=1\}}{\sum_{i=1}^n h(X_i,Y_i)\mathbf{1}\{A_i=1\}} - \frac{\sum_{i=1}^n g(\theta;X_i,Y_i)\mathbf{1}\{A_i=0\}}{\sum_{i=1}^n h(X_i,Y_i)\mathbf{1}\{A_i=0\}} = 0 \end{array} \right\}.$$

The corresponding empirical problem given the proxy sensitive attribute is

$$(\widetilde{\mathcal{P}}_n): \quad \widetilde{\theta}_n \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \frac{1}{n}\sum_{i=1}^n \ell(\theta; X_i, Y_i) \\ \text{subject to} & \frac{\sum_{i=1}^n g(\theta;X_i,Y_i)\mathbf{1}\{Z_i=1\}}{\sum_{i=1}^n h(X_i,Y_i)\mathbf{1}\{Z_i=1\}} - \frac{\sum_{i=1}^n g(\theta;X_i,Y_i)\mathbf{1}\{Z_i=0\}}{\sum_{i=1}^n h(X_i,Y_i)\mathbf{1}\{Z_i=0\}} = 0 \end{array} \right\}.$$

We denote

$$F(\theta) = \mathbb{E}\big[\ell(\theta; X, Y)\big], \widehat{F}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(\theta; X_i, Y_i),$$

$$G(\theta) = \frac{\mathbb{E}\big[g(\theta; X, Y)|A=1\big]}{\mathbb{E}\big[h(X, Y)|A=1\big]} - \frac{\mathbb{E}\big[g(\theta; X, Y)|A=0\big]}{\mathbb{E}\big[h(X, Y)|A=0\big]}$$

and

$$\widehat{G}_n(\theta) = \frac{\sum_{i=1}^n g(\theta; X_i, Y_i)\mathbf{1}\{A_i=1\}}{\sum_{i=1}^n h(X_i, Y_i)\mathbf{1}\{A_i=1\}} - \frac{\sum_{i=1}^n g(\theta; X_i, Y_i)\mathbf{1}\{A_i=0\}}{\sum_{i=1}^n h(X_i, Y_i)\mathbf{1}\{A_i=0\}}.$$

Note that $\widehat{F}_n(\cdot)$ and $\widehat{G}_n(\cdot)$'s are random functions serving as approximations to $F(\cdot)$ and $G(\cdot)$'s. Consider the Lagrangian functions

$$L(\theta, \lambda) = F(\theta) + \lambda G(\theta) \quad \text{and} \quad \widehat{L}_n(\theta, \lambda) = \widehat{F}_n(\theta) + \lambda \widehat{G}_n(\theta)$$

of the programs $(\mathcal{P}_0)$ and $(\mathcal{P}_n)$ respectively.

Lemma B.3.1 (A version of Theorem 6.6.2 in Rubinstein and Shapiro (1993))**.** *Suppose that:*

(i) The functions $F(\theta)$ and $G(\theta)$ are twice continuously differentiable.

(ii) *The true program $(\mathcal{P}_0)$ has a unique optimal solution $\theta^\star$ and a unique Lagrange multiplier $\lambda^\star$ with $\theta^\star$ being an interior point of Θ.*

(iii) *The Hessian matrix $\nabla^2 L(\theta^\star, \lambda^\star)$ is positive definite.*

(iv) *The random functions $\widehat{G}_n(\theta)$,$k \in [K]$, are Lipschitz continuous in a neighborhood of $\theta^\star$ and differentiable at $\theta^\star$ with probability 1.*

(v)

$$\|\Delta_{in}(\theta^\star)\|_2 = O_p(n^{-1/2}), \quad i = 1, 2, 3$$

and there is a neighborhood U of $\theta^\star$ such that

$$\sup_{\theta \in U} \frac{\|\Delta_{in}(\theta) - \Delta_{in}(\theta^\star)\|_2}{n^{-1/2} + \|\theta - \theta^\star\|_2} = o_p(1), \quad i = 1, 2, 3.$$

Here we define random mappings $\Delta_{1n}(\theta) = \nabla \widehat{F}_n(\theta) - \nabla F(\theta)$, $\Delta_{2n}(\theta) = \widehat{G}_n(\theta) - G(\theta)$, and $\Delta_{3n}(\theta) = \nabla \widehat{G}_n(\theta) - \nabla G(\theta)$.

(vi) *Random vectors $\sqrt{n}(\nabla \widehat{L}_n(\theta^\star, \lambda^\star), \widehat{G}_n(\theta^\star))$ converge in distribution to $\boldsymbol{Y} = (\boldsymbol{Y}_1, Y_2)$ as $n \to \infty$, where $\boldsymbol{Y}_1$ is a random vector and Y_2 is a random variable.*

Let $\widehat{\theta}_n$ be an optimal solution of $(\mathcal{P}_n)$ converging in probability as $n \to \infty$ to $\theta^\star$. Then

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{d} \bar{\boldsymbol{x}}(\boldsymbol{Y})$$

where $\bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}(\boldsymbol{Y})$ is the optimal solution to the quadratic programming problem

$$\begin{array}{ll} \underset{\boldsymbol{x}}{\text{minimize}} & \boldsymbol{x}^\top \boldsymbol{Y}_1 + \frac{1}{2} \boldsymbol{x}^\top \nabla^2 L(\theta^\star, \lambda^\star) \boldsymbol{x} \\ \text{subject to} & \nabla G(\theta^\star)^\top \boldsymbol{x} + Y_2 = 0 \end{array}.$$

Recall the standing assumptions, *(i)*, *(iv)*, *(v)* are guaranteed by the smoothness and concentration assumption, *(ii)* is postulated by the uniqueness assumption, and

(iii) is made by our assumption. Now we derive the limiting distribution of random vectors $\sqrt{n}(\nabla \widehat{L}_n(\theta^\star, \lambda^\star), \widehat{G}_n(\theta^\star))$ required in *(vi)*.

For $a \in \{0, 1\}$, we have

$$\mathbb{E}\big[g(\theta^\star; X, Y)\mathbf{1}\{A = a\}\big] = \mathbb{P}(A = a)\mathbb{E}\big[g(\theta^\star; X, Y)|A = a\big] = \pi_a \mathbb{E}_{Q_a}[g],$$

and

$$\begin{aligned}
\mathrm{Var}[g(\theta^\star; X, Y)\mathbf{1}\{A = a\}] &= \mathbb{E}\big[g^2(\theta^\star; X, Y)\mathbf{1}\{A = a\}\big] - \big\{\mathbb{E}\big[g(\theta^\star; X, Y)\mathbf{1}\{A = a\}\big]\big\}^2 \\
&= \pi_a \mathbb{E}_{Q_a}[g^2] - \pi_a^2 (\mathbb{E}_{Q_a}[g])^2 \\
&= \pi_a (\mathbb{E}_{Q_a}[g^2] - (\mathbb{E}_{Q_a}[g])^2) + (\pi_a - \pi_a^2)(\mathbb{E}_{Q_a}[g])^2 \\
&= \pi_a \,\mathrm{Var}_{Q_a}[g] + \pi_0 \pi_1 (\mathbb{E}_{Q_a}[g])^2.
\end{aligned}$$

Similarly, for $a \in \{0, 1\}$, we have $\mathbb{E}\big[h(X, Y)\mathbf{1}\{A = a\}\big] = \pi_a \mathbb{E}_{Q_a}[h]$ and

$$\mathrm{Var}[h(X, Y)\mathbf{1}\{A = a\}] = \pi_a \,\mathrm{Var}_{Q_a}[h] + \pi_0 \pi_1 (\mathbb{E}_{Q_a}[h])^2.$$

Moreover, we have

$$\begin{aligned}
&\mathrm{Cov}(g(\theta^\star; X, Y)\mathbf{1}\{A = 1\}, g(\theta^\star; X, Y)\mathbf{1}\{A = 0\}) \\
&= \mathbb{E}\big[g^2(\theta^\star; X, Y)\mathbf{1}\{A = 0\}\mathbf{1}\{A = 1\}\big] - \\
&\quad \mathbb{E}\big[g(\theta^\star; X, Y)\mathbf{1}\{A = 0\}\big] \times \mathbb{E}\big[g(\theta^\star; X, Y)\mathbf{1}\{A = 1\}\big] \\
&= -\pi_0 \pi_1 \mathbb{E}_{Q_0}[g]\mathbb{E}_{Q_1}[g]
\end{aligned}$$

and similarly we can derive

$$\mathrm{Cov}(h(X, Y)\mathbf{1}\{A = 1\}, h(X, Y)\mathbf{1}\{A = 0\}) = -\pi_0 \pi_1 \mathbb{E}_{Q_0}[h]\mathbb{E}_{Q_1}[h],$$

$$\mathrm{Cov}(g(\theta^\star;X,Y)\mathbf{1}\{A=a\}, h(X,Y)\mathbf{1}\{A=a\}) = \pi_a \,\mathrm{Cov}_{Q_a}[g,h] + \pi_0\pi_1 \mathbb{E}_{Q_a}[g]\mathbb{E}_{Q_a}[h]$$

and

$$\mathrm{Cov}(g(\theta^\star;X,Y)\mathbf{1}\{A=a\}, h(X,Y)\mathbf{1}\{A=1-a\}) = -\pi_0\pi_1 \mathbb{E}_{Q_a}[g]\mathbb{E}_{Q_{1-a}}[h]$$

for $a \in \{0,1\}$.

Let $\boldsymbol{\eta}_1 = \mathbb{E}\big[\nabla \ell(\theta^\star;X,Y)\big]$, $\boldsymbol{\eta}_2 = \pi_1 \mathbb{E}_{Q_1}\big[\nabla g(\theta^\star;X,Y)\big]$, $\boldsymbol{\eta}_3 = \pi_0 \mathbb{E}_{Q_0}\big[\nabla g(\theta^\star;X,Y)\big]$. By central limit theorem,

$$\sqrt{n}\left\{\begin{bmatrix} n^{-1}\sum_{i=1}^n \nabla \ell(\theta^\star;X_i,Y_i) \\ n^{-1}\sum_{i=1}^n \nabla g(\theta^\star;X_i,Y_i)\mathbf{1}\{A_i=1\} \\ n^{-1}\sum_{i=1}^n \nabla g(\theta^\star;X_i,Y_i)\mathbf{1}\{A_i=0\} \\ n^{-1}\sum_{i=1}^n g(\theta^\star;X_i,Y_i)\mathbf{1}\{A_i=1\} \\ n^{-1}\sum_{i=1}^n g(\theta^\star;X_i,Y_i)\mathbf{1}\{A_i=0\} \\ n^{-1}\sum_{i=1}^n h(X_i,Y_i)\mathbf{1}\{A_i=1\} \\ n^{-1}\sum_{i=1}^n h(X_i,Y_i)\mathbf{1}\{A_i=0\} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \\ \boldsymbol{\eta}_3 \\ \pi_1\mathbb{E}_{Q_1}[g] \\ \pi_0\mathbb{E}_{Q_0}[g] \\ \pi_1\mathbb{E}_{Q_1}[h] \\ \pi_0\mathbb{E}_{Q_0}[h] \end{bmatrix}\right\} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix}\right), \tag{B.3.1}$$

where $\Omega_{11} \in \mathbb{R}^{3d\times 3d}, \Omega_{21} \in \mathbb{R}^{4\times 3d}, \Omega_{12} = \Omega_{21}^\top$, Ω_{22} is given by $[\Omega_{22,\text{left}} \mid \Omega_{22,\text{right}}]$, where

$$\Omega_{22,\text{left}} = \begin{bmatrix} \pi_1 Q_1^2[g] + \pi_0\pi_1 (Q_1 g)^2 & -\pi_0\pi_1 Q_0 g Q_1 g \\ -\pi_0\pi_1 Q_0 g Q_1 g & \pi_0 Q_0^2[g] + \pi_0\pi_1 (Q_0 g)^2 \\ \pi_1 Q_1^2[g,h] + \pi_0\pi_1 Q_1 g Q_1 h & -\pi_0\pi_1 Q_0 g Q_1 h \\ -\pi_0\pi_1 Q_0 h Q_1 g & \pi_0 Q_0^2[g,h] + \pi_0\pi_1 Q_0 g Q_0 h \end{bmatrix}.$$

and

$$\Omega_{22,\text{right}} = \begin{bmatrix} \pi_1 Q_1^2[g,h] + \pi_0\pi_1 Q_1 g Q_1 h & -\pi_0\pi_1 Q_0 h Q_1 g \\ -\pi_0\pi_1 Q_0 g Q_1 h & \pi_0 Q_0^2[g,h] + \pi_0\pi_1 Q_0 g Q_0 h \\ \pi_1 Q_1^2[h] + \pi_0\pi_1 (Q_1 h)^2 & -\pi_0\pi_1 Q_0 h Q_1 h \\ -\pi_0\pi_1 Q_0 h Q_1 h & \pi_0 Q_0^2[h] + \pi_0\pi_1 (Q_0 h)^2 \end{bmatrix}.$$

Let function $w : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}^{d+1}$ be

$$w(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, s_1, s_2, s_3, s_4) = \left(\mathbf{v}_1 + \lambda^\star \left\{ \frac{\mathbf{v}_2}{s_3} - \frac{\mathbf{v}_3}{s_4} \right\}, \frac{s_1}{s_3} - \frac{s_2}{s_4} \right)^\top.$$

The gradient of function w evaluated at

$$(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \boldsymbol{\eta}_3)$$

and

$$(s_1, s_2, s_3, s_4) = (\pi_1 \mathbb{E}_{Q_1}[g], \pi_0 \mathbb{E}_{Q_0}[g], \pi_1 \mathbb{E}_{Q_1}[h], \pi_0 \mathbb{E}_{Q_0}[h])$$

is given by

$$\nabla w = \begin{bmatrix} *_{3d \times d} & \mathbf{0}_{3d \times 1} \\ *_{4 \times d} & \xi_{4 \times 1} \end{bmatrix} \in \mathbb{R}^{(3d+4) \times (d+1)}$$

where

$$\xi = \left(\frac{1}{\pi_1 Q_1 h}, -\frac{1}{\pi_0 Q_0 h}, -\frac{Q_1 g}{\pi_1 (Q_1 h)^2}, \frac{Q_0 g}{\pi_0 (Q_0 h)^2} \right)^\top.$$

Applying delta method to (B.3.1) with $w(\cdot)$, we have

$$\sqrt{n} \begin{bmatrix} \nabla \widehat{L}_n(\theta^\star, \lambda^\star) \\ \widehat{G}_n(\theta^\star) \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \nabla w^\top \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix} \nabla w \right) \overset{d}{=\!=} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \sigma^2 \end{bmatrix} \right),$$

where

$$\begin{aligned} \sigma^2 = \xi^\top \Omega_{22} \xi = & \frac{Q_0^2[g]}{\pi_0 (Q_0 h)^3} + \frac{Q_0^2[h](Q_0 g)^2}{\pi_0 (Q_0 h)^4} - \frac{2 Q_0^2[g,h] Q_0 g}{\pi_0 (Q_0 h)^3} + \\ & \frac{Q_1^2[g]}{\pi_1 (Q_1 h)^3} + \frac{Q_1^2[h](Q_1 g)^2}{\pi_1 (Q_1 h)^4} - \frac{2 Q_1^2[g,h] Q_1 g}{\pi_1 (Q_1 h)^3} \end{aligned} \tag{B.3.2}$$

Note that KKT condition implies

$$\boldsymbol{\eta}_1 + \lambda^\star \left\{ \frac{\boldsymbol{\eta}_2}{\pi_1 Q_1 g} - \frac{\boldsymbol{\eta}_3}{\pi_0 Q_0 g} \right\} = 0 \quad \text{and} \quad \frac{Q_1 g}{Q_1 h} = \frac{Q_0 g}{Q_0 h} \triangleq \kappa. \tag{B.3.3}$$

Combining (B.3.2) and (B.3.3), we have

$$\begin{aligned}
&\sigma^2 \\
=&\frac{\mathrm{Var}_{Q_0}[g] + \mathrm{Var}_{Q_0}[\kappa h] - 2\,\mathrm{Cov}_{Q_0}[g, \kappa h]}{\pi_0(\mathbb{E}_{Q_0}[h])^2} + \frac{\mathrm{Var}_{Q_1}[g] + \mathrm{Var}_{Q_1}[\kappa h] - 2\,\mathrm{Cov}_{Q_1}[g, \kappa h]}{\pi_1(\mathbb{E}_{Q_1}[h])^2} \\
=&\frac{\mathrm{Var}_{Q_0}[g - \kappa h]}{\pi_0(\mathbb{E}_{Q_0}[h])^2} + \frac{\mathrm{Var}_{Q_1}[g - \kappa h]}{\pi_1(\mathbb{E}_{Q_1}[h])^2}.
\end{aligned} \tag{B.3.4}$$

Therefore, we conclude that the limiting distribution of $\sqrt{n}(\nabla \widehat{L}_n(\theta^\star, \lambda^\star), G_n(\theta^\star))$ is

$$(\boldsymbol{Y}_1, Y_2) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \sigma^2 \end{bmatrix}\right).$$

By Lemma (B.3.1), we have

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{d} \bar{\boldsymbol{x}},$$

where $\bar{\boldsymbol{x}}$ is given by the linear system

$$\underbrace{\begin{bmatrix} \nabla^2 L(\theta^\star, \lambda^\star) & \nabla G(\theta^\star) \\ \nabla G(\theta^\star)^\top & 0 \end{bmatrix}}_{\triangleq B} \begin{bmatrix} \bar{\boldsymbol{x}} \\ \bar{\lambda} \end{bmatrix} = - \begin{bmatrix} \boldsymbol{Y}_1 \\ Y_2 \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \sigma^2 \end{bmatrix}\right),$$

or

$$\begin{bmatrix} \bar{\boldsymbol{x}} \\ \bar{\lambda} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, B^{-1} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \sigma^2 \end{bmatrix} B^{-1}\right), \tag{B.3.5}$$

which implies $\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{d} \bar{\boldsymbol{x}} \sim \mathcal{N}(\mathbf{0}, \bar{\Sigma})$ for some $\bar{\mu}$ and $\bar{\Sigma}$ determined by (B.3.5).

By delta method, we have

$$\sqrt{n} G(\widehat{\theta}_n) = \sqrt{n}\{G(\widehat{\theta}_n) - \underbrace{G(\theta^\star)}_{=0}\} \xrightarrow{d} \mathcal{N}(0, \nabla G(\theta^\star)^\top \bar{\Sigma} \nabla G(\theta^\star)).$$

Now we calculate $\nabla G(\theta^\star)^\top \bar{\Sigma} \nabla G(\theta^\star)$.

For notation simplicity, we denote $\nabla^2 L = \nabla^2 L(\theta^\star, \lambda^\star), \nabla G = \nabla G(\theta^\star)$ and $H = (\nabla^2 L)^{-1} \nabla G [\nabla G^\top (\nabla^2 L)^{-1} \nabla G]^{-1}$. By block matrix inversion, we have

$$B^{-1} = \begin{bmatrix} (\nabla^2 L)^{-1} - H \nabla G^\top (\nabla^2 L)^{-1} & H \\ H^\top & -[\nabla G^\top (\nabla^2 L)^{-1} \nabla G]^{-1}. \end{bmatrix}$$

Note that $\nabla G^\top H = 1$ and $\nabla G^\top \left\{ (\nabla^2 L)^{-1} - H \nabla G^\top (\nabla^2 L)^{-1} \right\} = 0$. We have

$$\begin{aligned}
& \nabla G(\theta^\star)^\top \bar{\Sigma} \nabla G(\theta^\star) \\
= & \nabla G^\top \left[\left\{ (\nabla^2 L)^{-1} - H \nabla G^\top (\nabla^2 L)^{-1} \right\} \Sigma_{11} + H \Sigma_{21} \right] \\
& \underbrace{\left\{ (\nabla^2 L)^{-1} - (\nabla^2 L)^{-1} \nabla G H^\top \right\} \nabla G}_{=0} + \\
& \underbrace{\nabla G^\top \left\{ (\nabla^2 L)^{-1} - H \nabla G^\top (\nabla^2 L)^{-1} \right\}}_{=0} \Sigma_{12} H^\top \nabla G + \nabla G^\top H \sigma^2 H^\top \nabla G \\
= & \sigma^2.
\end{aligned}$$

Therefore, we conclude that

$$\begin{aligned}
\sqrt{n} \{ c(\widehat{\theta}_n) - \cancel{c(\theta^\star)} \} &= \sqrt{n} G(\widehat{\theta}_n) \\
&\xrightarrow{d} \mathcal{N}(0, \sigma^2) \overset{d}{=} \mathcal{N}\left(0, \frac{\mathrm{Var}_{Q_0}[g - \kappa h]}{\pi_0 (\mathbb{E}_{Q_0}[h])^2} + \frac{\mathrm{Var}_{Q_1}[g - \kappa h]}{\pi_1 (\mathbb{E}_{Q_1}[h])^2} \right).
\end{aligned}$$

By a similar argument, we have

$$\sqrt{n} \{ \psi_{\mathrm{frac}} \times c(\widetilde{\theta}_n) - \cancel{\psi_{\mathrm{frac}} \times c(\theta^\star)} \} \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathrm{Var}_{\widetilde{Q}_0}[g - \kappa h]}{\widetilde{\pi}_0 (\mathbb{E}_{\widetilde{Q}_0}[h])^2} + \frac{\mathrm{Var}_{\widetilde{Q}_1}[g - \kappa h]}{\widetilde{\pi}_1 (\mathbb{E}_{\widetilde{Q}_1}[h])^2} \right),$$

which implies

$$\sqrt{n} \times c(\widetilde{\theta}_n) \xrightarrow{d} \mathcal{N}(0, \widetilde{\sigma}^2) \overset{d}{=} \mathcal{N}\left(0, \psi_{\mathrm{frac}}^{-2} \times \left\{ \frac{\mathrm{Var}_{\widetilde{Q}_0}[g - \kappa h]}{\widetilde{\pi}_0 (\mathbb{E}_{\widetilde{Q}_0}[h])^2} + \frac{\mathrm{Var}_{\widetilde{Q}_1}[g - \kappa h]}{\widetilde{\pi}_1 (\mathbb{E}_{\widetilde{Q}_1}[h])^2} \right\} \right).$$

Now, we prove the case when $\alpha_n = o(\frac{1}{\sqrt{n}})$. For this case note that the equality constraint for the population problem can be rewritten as two inequality constraints:

$$(\mathcal{P}_0): \quad \theta^\star \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \mathbb{E}\big[\ell(\theta; X, Y)\big] \\ \text{subject to} & \dfrac{\mathbb{E}\big[g(\theta; X, Y)|A=1\big]}{\mathbb{E}\big[h(X,Y)|A=1\big]} - \dfrac{\mathbb{E}\big[g(\theta; X, Y)|A=0\big]}{\mathbb{E}\big[h(X,Y)|A=0\big]} \leq 0 \\ & \dfrac{\mathbb{E}\big[g(\theta; X, Y)|A=0\big]}{\mathbb{E}\big[h(X,Y)|A=0\big]} - \dfrac{\mathbb{E}\big[g(\theta; X, Y)|A=1\big]}{\mathbb{E}\big[h(X,Y)|A=1\big]} \leq 0 \end{array} \right\},$$

where the expectation is with respect to the underlying distribution of tuple (X, A, Y).

The corresponding empirical problem given the true sensitive attribute is

$$(\mathcal{P}_n): \quad \widehat{\theta}_n \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \frac{1}{n}\sum_{i=1}^n \ell(\theta; X_i, Y_i) \\ \text{subject to} & \frac{\sum_{i=1}^n g(\theta;X_i,Y_i)\mathbf{1}\{A_i=1\}}{\sum_{i=1}^n h(X_i,Y_i)\mathbf{1}\{A_i=1\}} - \frac{\sum_{i=1}^n g(\theta;X_i,Y_i)\mathbf{1}\{A_i=0\}}{\sum_{i=1}^n h(X_i,Y_i)\mathbf{1}\{A_i=0\}} - \alpha_n \leq 0 \\ & \frac{\sum_{i=1}^n g(\theta;X_i,Y_i)\mathbf{1}\{A_i=0\}}{\sum_{i=1}^n h(X_i,Y_i)\mathbf{1}\{A_i=0\}} - \frac{\sum_{i=1}^n g(\theta;X_i,Y_i)\mathbf{1}\{A_i=1\}}{\sum_{i=1}^n h(X_i,Y_i)\mathbf{1}\{A_i=1\}} - \alpha_n \leq 0 \end{array} \right\}.$$

The corresponding empirical problem given the proxy sensitive attribute is

$$(\widetilde{\mathcal{P}}_n): \quad \widetilde{\theta}_n \in \left\{ \begin{array}{ll} \arg\min_{\theta\in\Theta} & \frac{1}{n}\sum_{i=1}^n \ell(\theta; X_i, Y_i) \\ \text{subject to} & \frac{\sum_{i=1}^n g(\theta;X_i,Y_i)\mathbf{1}\{Z_i=1\}}{\sum_{i=1}^n h(X_i,Y_i)\mathbf{1}\{Z_i=1\}} - \frac{\sum_{i=1}^n g(\theta;X_i,Y_i)\mathbf{1}\{Z_i=0\}}{\sum_{i=1}^n h(X_i,Y_i)\mathbf{1}\{Z_i=0\}} - \alpha_n \leq 0 \\ & \frac{\sum_{i=1}^n g(\theta;X_i,Y_i)\mathbf{1}\{Z_i=0\}}{\sum_{i=1}^n h(X_i,Y_i)\mathbf{1}\{Z_i=0\}} - \frac{\sum_{i=1}^n g(\theta;X_i,Y_i)\mathbf{1}\{Z_i=1\}}{\sum_{i=1}^n h(X_i,Y_i)\mathbf{1}\{Z_i=1\}} - \alpha_n \leq 0 \end{array} \right\}.$$

We denote

$$F(\theta) = \mathbb{E}\big[\ell(\theta; X, Y)\big], \widehat{F}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(\theta; X_i, Y_i),$$

$$G_1(\theta) = \frac{\mathbb{E}\big[g(\theta; X, Y)|A=1\big]}{\mathbb{E}\big[h(X,Y)|A=1\big]} - \frac{\mathbb{E}\big[g(\theta; X, Y)|A=0\big]}{\mathbb{E}\big[h(X,Y)|A=0\big]},$$

$$G_2(\theta) = \frac{\mathbb{E}\big[g(\theta; X, Y)|A=0\big]}{\mathbb{E}\big[h(X,Y)|A=0\big]} - \frac{\mathbb{E}\big[g(\theta; X, Y)|A=1\big]}{\mathbb{E}\big[h(X,Y)|A=1\big]},$$

$$\widehat{G}_{1n}(\theta) = \frac{\sum_{i=1}^n g(\theta; X_i, Y_i)\mathbf{1}\{A_i = 1\}}{\sum_{i=1}^n h(X_i, Y_i)\mathbf{1}\{A_i = 1\}} - \frac{\sum_{i=1}^n g(\theta; X_i, Y_i)\mathbf{1}\{A_i = 0\}}{\sum_{i=1}^n h(X_i, Y_i)\mathbf{1}\{A_i = 0\}} - \alpha_n,$$

and

$$\widehat{G}_{2n}(\theta) = \frac{\sum_{i=1}^n g(\theta; X_i, Y_i)\mathbf{1}\{A_i = 0\}}{\sum_{i=1}^n h(X_i, Y_i)\mathbf{1}\{A_i = 0\}} - \frac{\sum_{i=1}^n g(\theta; X_i, Y_i)\mathbf{1}\{A_i = 1\}}{\sum_{i=1}^n h(X_i, Y_i)\mathbf{1}\{A_i = 1\}} - \alpha_n.$$

Consider the Lagrangian functions

$$L(\theta, \boldsymbol{\lambda}) = F(\theta) + \lambda_1 G_1(\theta) + \lambda_2 G_2(\theta) \quad \text{and} \quad \widehat{L}_n(\theta, \boldsymbol{\lambda}) = \widehat{F}_n(\theta) + \lambda_1 \widehat{G}_{1n}(\theta) + \lambda_2 \widehat{G}_{2n}(\theta).$$

of the programs $(\mathcal{P}_0)$ and $(\mathcal{P}_n)$ respectively.

Lemma B.3.2 (A version of Theorem 6.6.2 in Rubinstein and Shapiro (1993))**.** *Suppose that:*

(i) The functions $F(\theta)$, $G_1(\theta)$ and $G_2(\theta)$ are twice continuously differentiable.

(ii) The true program $(\mathcal{P}_0)$ has a unique optimal solution $\theta^\star$ and a unique Lagrange multiplier $\boldsymbol{\lambda}^\star$ with $\theta^\star$ being an interior point of Θ.

(iii) The Hessian matrix $\nabla^2 L(\theta^\star, \boldsymbol{\lambda}^\star)$ is positive definite.

(iv) The random functions $\widehat{G}_{1n}(\theta)$ and $\widehat{G}_{2n}(\theta)$,$k \in [K]$, are Lipschitz continuous in a neighborhood of $\theta^\star$ and differentiable at $\theta^\star$ with probability 1.

(v)

$$\|\Delta_{in}(\theta^\star)\|_2 = O_p(n^{-1/2}), \quad i = 1, 2, 3$$

and there is a neighborhood U of $\theta^\star$ such that

$$\sup_{\theta \in U} \frac{\|\Delta_{in}(\theta) - \Delta_{in}(\theta^\star)\|_2}{n^{-1/2} + \|\theta - \theta^\star\|_2} = o_p(1), \quad i = 1, 2, 3.$$

Here we define random mappings $\Delta_{1n}(\theta) = \nabla \widehat{F}_n(\theta) - \nabla F(\theta)$, $\Delta_{2n}(\theta) = \widehat{G}_n(\theta) - G(\theta)$, *and* $\Delta_{3n}(\theta) = \nabla \widehat{G}_n(\theta) - \nabla G(\theta)$.

(vi) Random vectors $\sqrt{n}(\nabla \widehat{L}_n(\theta^\star, \boldsymbol{\lambda}^\star), \widehat{G}_{1n}(\theta^\star)), \widehat{G}_{2n}(\theta^\star))$ *converge in distribution to* $\boldsymbol{Y} = (\boldsymbol{Y}_1, Y_2, Y_3)$ *as* $n \to \infty$, *where* $\boldsymbol{Y}_1$ *is a random vector and* Y_2 *and* Y_3 *are random variables.*

Let $\widehat{\theta}_n$ *be an optimal solution of* $(\mathcal{P}_n)$ *converging in probability as* $n \to \infty$ *to* $\theta^\star$. *Then*

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{d} \bar{\boldsymbol{x}}(\boldsymbol{Y})$$

where $\bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}(\boldsymbol{Y})$ *is the optimal solution to the quadratic programming problem*

$$\begin{array}{ll} \underset{\boldsymbol{x}}{\text{minimize}} & \boldsymbol{x}^\top \boldsymbol{Y}_1 + \frac{1}{2}\boldsymbol{x}^\top \nabla^2 L(\theta^\star, \boldsymbol{\lambda}^\star)\boldsymbol{x} \\ \text{subject to} & \nabla G_1(\theta^\star)^\top \boldsymbol{x} + Y_2 \leq 0 \\ & \nabla G_2(\theta^\star)^\top \boldsymbol{x} + Y_3 \leq 0 \end{array} .$$

Note that

$$\nabla G_1(\theta^\star)^\top \boldsymbol{x} + Y_2 \leq 0 \iff \nabla G(\theta^\star)^\top \boldsymbol{x} + Y \leq 0$$

and

$$\nabla G_1(\theta^\star)^\top \boldsymbol{x} + Y_2 \leq 0 \iff -\nabla G(\theta^\star)^\top \boldsymbol{x} + (-Y) \leq 0.$$

Therefore the last quadratic programming problem with two inequality constraints reduces to the quadratic programming problem with single equality constraint when $\alpha_n \equiv 0$. The limiting distributional results thus persist as we proved for the $\alpha_n \equiv 0$ case.

Lastly, we calculate the asymptotic relative efficiency (ARE) of $\widetilde{\theta}_n$ to $\widehat{\theta}_n$.

Recall that

$$\begin{aligned}
\sigma^2 &= \frac{\mathrm{Var}_{Q_0}[g-\kappa h]}{\pi_0(\mathbb{E}_{Q_0}[h])^2} + \frac{\mathrm{Var}_{Q_1}[g-\kappa h]}{\pi_1(\mathbb{E}_{Q_1}[h])^2}, \\
\widetilde{\sigma}^2 &= \psi_{\mathrm{frac}}^{-2} \times \left\{ \frac{\mathrm{Var}_{\widetilde{Q}_0}[g-\kappa h]}{\widetilde{\pi}_0(\mathbb{E}_{\widetilde{Q}_0}[h])^2} + \frac{\mathrm{Var}_{\widetilde{Q}_1}[g-\kappa h]}{\widetilde{\pi}_1(\mathbb{E}_{\widetilde{Q}_1}[h])^2} \right\} \\
&= \psi_{\mathrm{frac}}^{-2} \times \left\{ \frac{(1-\gamma)\pi_0 \,\mathrm{Var}_{Q_0}[g-\kappa h] + \gamma\pi_1 \,\mathrm{Var}_{Q_1}[g-\kappa h]}{\{(1-\gamma)\pi_0\mathbb{E}_{Q_0}[h] + \gamma\pi_1\mathbb{E}_{Q_1}[h]\}^2} \right. \\
&\qquad \left. + \frac{\gamma\pi_0 \,\mathrm{Var}_{Q_0}[g-\kappa h] + (1-\gamma)\pi_1 \,\mathrm{Var}_{Q_1}[g-\kappa h]}{\{\gamma\pi_0\mathbb{E}_{Q_0}[h] + (1-\gamma)\pi_1\mathbb{E}_{Q_1}[h]\}^2} \right\},
\end{aligned}$$

and

$$\psi_{\mathrm{frac}} = \frac{(1-2\gamma)\pi_0\pi_1\mathbb{E}_{Q_0}[h]\mathbb{E}_{Q_1}[h]}{\{\gamma\pi_0\mathbb{E}_{Q_0}[h] + (1-\gamma)\pi_1\mathbb{E}_{Q_1}[h]\}\{(1-\gamma)\pi_0\mathbb{E}_{Q_0}[h] + \gamma\pi_1\mathbb{E}_{Q_1}[h]\}}.$$

Therefore, we have

$$\begin{aligned}
\mathrm{ARE}(\widetilde{\theta}_n, \widehat{\theta}_n) = \frac{\sigma^2}{\widetilde{\sigma}^2} &= \varphi\left(\gamma, \frac{\pi_0\mathbb{E}_{Q_0}[h]}{\pi_1\mathbb{E}_{Q_1}[h]}, \frac{\mathrm{Var}_{Q_0}[g(\theta^\star;X,Y)-\kappa h(X,Y)]/\mathbb{E}_{Q_0}[h]}{\mathrm{Var}_{Q_1}[g(\theta^\star;X,Y)-\kappa h(X,Y)]/\mathbb{E}_{Q_1}[h]}\right) \\
&= \varphi\left(\gamma, \frac{\pi_0 m_0}{\pi_1 m_1}, \frac{\mathrm{Var}[g(\theta^\star;X,Y)-\kappa h(X,Y)|A=0]/m_0}{\mathrm{Var}[g(\theta^\star;X,Y)-\kappa h(X,Y)|A=1]/m_1}\right),
\end{aligned}$$

where

$$\varphi(\gamma, r_1, r_2) \triangleq \frac{(1-2\gamma)^2 r_1(r_1+r_2)}{\{\gamma r_1 + (1-\gamma)\}^2\{(1-\gamma)r_1r_2+\gamma\} + \{(1-\gamma)r_1+\gamma\}^2\{\gamma r_1 r_2 + (1-\gamma)\}}.$$

Hence we complete the proof of Theorem 3.3.2. □

B.4 Multiple Demographic Groups

We provide further discussion to supplement Section 3.3.3.

Note that the fairness notion (3.3.12) uses group 0 as a reference group. One can

also define a fairness notion by

$$\frac{\mathbb{E}\big[g(\theta;X,Y)|A=k\big]}{\mathbb{E}\big[h(X,Y)|A=k\big]}-\frac{\mathbb{E}\big[g(\theta;X,Y)\big]}{\mathbb{E}\big[h(X,Y)\big]}=0\quad\text{for } k\in\{0\}\cup[K]$$

which is symmetric in group indices. Due to the equivalence of (3.3.13) and (3.3.12), we opt to use (3.3.12) for a comparison with two-group theory.

Theorem 3.3.4 is a direct extension of Theorem 3.3.2 and follows the same proof procedure as of Theorem 3.3.2. Moreover, let $h\equiv 1$, the linear-fractional fairness (3.3.12) degenerates into linear fairness:

$$\mathbb{E}\big[g(\theta;X,Y)|A=k\big]-\mathbb{E}\big[g(\theta;X,Y)|A=0\big]=0\quad\text{for } k\in[K]. \tag{B.4.1}$$

By Theorem 3.3.4, we immediately have the following corollary.

Theorem B.4.1 (Privacy cost in linear fairness (B.4.1)-aware learning)**.** *Under the standing assumptions, let estimators $\widehat{\theta}_n$ and $\widetilde{\theta}_n$ be consistent for $\theta^\star$, then*

$$\sqrt{n}\{\boldsymbol{c}(\widehat{\theta}_n)-\cancel{\boldsymbol{c}(\theta^\star)}\}\xrightarrow{d}\mathcal{N}(\mathbf{0},\Sigma)\ \textit{and}\ \sqrt{n}\{\boldsymbol{c}(\widetilde{\theta}_n)-\cancel{\boldsymbol{c}(\theta^\star)}\}\xrightarrow{d}\mathcal{N}(\mathbf{0},\Psi_{\text{lin}}^{-1}\widetilde{\Sigma}\Psi_{\text{lin}}^{-\top}),$$

where

$$\Sigma_{kl}=\frac{\text{Var}_{Q_0}[g(\theta^\star;X,Y)]}{\pi_0}+\left(\frac{\text{Var}_{Q_k}[g(\theta^\star;X,Y)]}{\pi_k}\right)\mathbf{1}\{k=l\}$$

$$\widetilde{\Sigma}_{kl}=\frac{\text{Var}_{\widetilde{Q}_0}[g(\theta^\star;X,Y)]}{\widetilde{\pi}_0}+\left(\frac{\text{Var}_{\widetilde{Q}_k}[g(\theta^\star;X,Y)]}{\widetilde{\pi}_k}\right)\mathbf{1}\{k=l\}$$

$$\Psi_{\text{lin}}=\begin{cases}\left(\frac{1-K\gamma}{\widetilde{\pi}_k}-\frac{\gamma}{\widetilde{\pi}_0}\right)\pi_k & \textit{if } k=l\\ \left(\frac{1}{\widetilde{\pi}_k}-\frac{1}{\widetilde{\pi}_0}\right)\gamma\pi_l & \textit{if } k\neq l\end{cases}$$

for $k,l\in[K]$.

B.5 Missing Sensitive Attributes

Under the missingness mechanism (3.3.8), the probability of observing a complete sample from group a is

$$\mathbb{P}(A = a, R = 1) = \omega_a \pi_a$$

for $a \in \{0, 1\}$. By the intermediate conclusion of Theorem 3.3.2, we have

$$\sqrt{n}\{c(\widehat{\theta}_n) - \cancel{c(\theta^\star)}\} \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathrm{Var}_{Q_0}[g(\theta^\star; X, Y) - \kappa h(X, Y)]}{\pi_0(\mathbb{E}_{Q_0}[h(X, Y)])^2} + \frac{\mathrm{Var}_{Q_1}[g(\theta^\star; X, Y) - \kappa h(X, Y)]}{\pi_1(\mathbb{E}_{Q_1}[h(X, Y)])^2}\right),$$

and

$$\sqrt{n}\{c(\widetilde{\theta}_n) - \cancel{c(\theta^\star)}\} \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathrm{Var}_{Q_0}[g(\theta^\star; X, Y) - \kappa h(X, Y)]}{\omega_0\pi_0(\mathbb{E}_{Q_0}[h(X, Y)])^2} + \frac{\mathrm{Var}_{Q_1}[g(\theta^\star; X, Y) - \kappa h(X, Y)]}{\omega_1\pi_1(\mathbb{E}_{Q_1}[h(X, Y)])^2}\right).$$

Comparing the two asymptotic variances, we conclude that

$$\mathrm{ARE}(\widetilde{\theta}_n, \widehat{\theta}_n) = \frac{r_2 + r_1}{\omega_0^{-1} r_2 + \omega_1^{-1} r_1},$$

where

$$r_1 = \frac{\pi_0 m_0}{\pi_1 m_1} \text{ and } r_2 = \frac{\mathrm{Var}[g(\theta^\star; X, Y) - \kappa h(X, Y) | A = 0]/m_0}{\mathrm{Var}[g(\theta^\star; X, Y) - \kappa h(X, Y) | A = 1]/m_1}.$$

APPENDIX C

Appendix of Chapter IV

In Section C.1, the missing deriving steps for the toy example in Section 4.2.4 are provided. In Section C.2, all of the deferred proofs are presented. In Section C.3, we provide omitted experimental details and additional empirical results. In Section C.4, we generalize the KL divergence DR performative prediction framework to a general φ-divergence DR performative prediction version, and propose an algorithm to find the associated DRPO. In Section C.5, we relate some existing algorithmic convergence results for finding the PO to our algorithms.

C.1 Toy Example in Section 4.2.4 (Continued)

Recall that $\mathcal{Z} = \mathbb{R}$, $\Theta = [-1, 1]$, $\ell(z; \theta) = \theta z$, and the nominal distribution map is $\mathcal{D}(\theta) = \mathcal{N}(f(\theta), \sigma^2)$ for some $f : [-1, 1] \to \mathbb{R}$ and $\sigma^2 > 0$. Then the nominal performative risk is

$$\mathrm{PR}(\theta) = \mathbb{E}_{Z \sim \mathcal{N}(f(\theta), \sigma^2)}[\ell(Z; \theta)] = \theta f(\theta).$$

We firstly show that the distributionally robust performative risk is given by

$$\mathrm{DRPR}(\theta) = \mathrm{PR}(\theta) + \sqrt{\rho}\,\mathrm{Penalty}(\theta),$$

where the penalty function $\text{Penalty}(\theta) = \sqrt{2\sigma^2}\,|\theta|$.

Proof. By Proposition 4.3.1, the strong duality of DRPR, and the well-established moment generating function of Gaussian distribution, we have

$$\begin{aligned}
\text{DRPR}(\theta) &= \sup_{\widetilde{\mathcal{D}}:\widetilde{\mathcal{D}}\in\mathcal{U}(\mathcal{D})} \mathbb{E}_{Z\sim\widetilde{\mathcal{D}}(\theta)}[\ell(Z;\theta)] \\
&= \inf_{\mu\geq 0}\left\{\mu \log \mathbb{E}_{Z\sim\mathcal{D}(\theta)}\left[e^{\ell(Z;\theta)/\mu}\right] + \mu\rho\right\} \\
&= \inf_{\mu\geq 0}\left\{\mu\left[\frac{\theta f(\theta)}{\mu} + \frac{\sigma^2\theta^2}{2\mu^2}\right] + \mu\rho\right\} \\
&= \theta f(\theta) + \inf_{\mu\geq 0}\left\{\frac{\sigma^2\theta^2}{2\mu} + \mu\rho\right\} \\
&= \theta f(\theta) + \sqrt{2\rho\sigma^2}\,|\theta| = \text{PR}(\theta) + \sqrt{\rho}\,\text{Penalty}(\theta).
\end{aligned}$$

Therefore, we derive an alternative form of $\text{DRPR}(\theta)$, which is $\text{PR}(\theta)$ with an L_1-regularization term. □

Now we are in a more concrete setup that $f(\theta) = a_1\theta + a_0$ for some $a_1, a_0 > 0$. Recall that for any $\widetilde{\mathcal{D}} \in \mathcal{U}(\mathcal{D})$, we denote the performative risk of $\widetilde{\mathcal{D}}$ be

$$\text{PR}_{\widetilde{\mathcal{D}}}(\theta) = \mathbb{E}_{Z\sim\widetilde{\mathcal{D}}(\theta)}[\ell(Z;\theta)].$$

We secondly show that

$$\sup_{\widetilde{\mathcal{D}}\in\mathcal{U}(\mathcal{D})} \text{PR}_{\widetilde{\mathcal{D}}}(\theta_{\text{PO}}) \geq \sup_{\widetilde{\mathcal{D}}\in\mathcal{U}(\mathcal{D})} \text{PR}_{\widetilde{\mathcal{D}}}(\theta_{\text{DRPO}}) + \frac{\rho\sigma^2}{2a_1}$$

for any fixed $\rho \leq \frac{a_0^2}{2\sigma^2}$.

Proof. In fact, we have

$$\theta_{\text{PO}} = \frac{-a_0}{2a_1}, \theta_{\text{DRPO}} = \frac{\sqrt{2\rho\sigma^2} - a_0}{2a_1}, \text{and } \text{DRPR}(\theta_{\text{DRPO}}) = \frac{-(a_0 - \sqrt{2\rho\sigma^2})^2}{4a_1}.$$

By Proposition 4.2.1, the generalization principle of distributionally robust performative prediction, we have

$$\sup_{\widetilde{\mathcal{D}}\in\mathcal{U}(\mathcal{D})} \mathrm{PR}_{\widetilde{\mathcal{D}}}(\theta_{\mathrm{DRPO}}) \le \mathrm{DRPR}(\theta_{\mathrm{DRPO}}) = \frac{-(a_0 - \sqrt{2\rho\sigma^2})^2}{4a_1}.$$

On the other hand, consider an *adversarial distribution map* $\mathcal{D}_{\mathrm{adv}}(\theta) = \mathcal{N}(a_1\theta + a_0 - \sqrt{2\rho\sigma^2}, \sigma^2)$. One can show that $D_{\mathrm{KL}}(\mathcal{D}_{\mathrm{adv}}(\theta), \mathcal{D}(\theta)) = \rho$ for all $\theta \in \Theta$, and therefore $\mathcal{D}_{\mathrm{adv}} \in \mathcal{U}(\mathcal{D})$. Then

$$\sup_{\widetilde{\mathcal{D}}\in\mathcal{U}(\mathcal{D})} \mathrm{PR}_{\widetilde{\mathcal{D}}}(\theta_{\mathrm{PO}}) \ge \mathrm{PR}_{\mathcal{D}_{\mathrm{adv}}}(\theta_{\mathrm{PO}}) = \frac{-(a_0 - \sqrt{2\rho\sigma^2})^2}{4a_1} + \frac{\rho\sigma^2}{2a_1}.$$

Therefore we complete the proof. □

Multi-variate case. We extend the uni-variate example to multi-variate case. Let $\mathcal{Z} = \mathbb{R}^d$, $\Theta = [-1, 1]^d$, and $\ell(z; \theta) = \theta^\top z$. Let the distribution map be $\mathcal{D}(\theta) = \mathcal{N}_d(f(\theta), \Sigma)$ for some $f : [-1, 1]^d \to \mathbb{R}^d$ and Σ is positive semi-definite. One can show that the performative risk is

$$\mathrm{PR}(\theta) = \mathbb{E}_{Z\sim\mathcal{D}(\theta)}[\ell(Z; \theta)] = \mathbb{E}_{Z\sim\mathcal{N}_d(f(\theta),\Sigma)}[\ell(Z; \theta)] = \theta^\top f(\theta)$$

and the distributionally robust performative risk is

$$\mathrm{DRPR}(\theta) = \sup_{\widetilde{\mathcal{D}}:\widetilde{\mathcal{D}}\in\mathcal{U}(\mathcal{D})} \mathbb{E}_{Z\sim\widetilde{\mathcal{D}}(\theta)}[\ell(Z; \theta)] = \theta^\top f(\theta) + \sqrt{2\rho\theta^\top\Sigma\theta} = \mathrm{PR}(\theta) + \sqrt{\rho}\,\mathrm{Penalty}(\theta), \tag{C.1.1}$$

where the penalty function $\mathrm{Penalty}(\theta) = \sqrt{2\theta^\top\Sigma\theta}$ penalizes the deviation of θ from the origin 0, and the critical radius ρ tunes the level of regularization. We also note that in this case the distributionally robust performative prediction problem can be transformed to a *second-order cone constrained program* (C.1.1).

C.2 Deferred Proofs

For simplicity of notation, we denote the true PO by $\theta^\star_{\mathrm{PO}}$, which is denoted by $\theta_{\mathrm{PO,true}}$ in the main article.

C.2.1 Proof of Proposition 4.3.1 and Proposition 4.3.2

Proof of Proposition 4.3.1. We only have to show that for any fixed $\theta \in \Theta$, the dual form (4.3.1) holds. This follows Theorem 1 presented in Hu and Hong (2013). □

We refer the readers to Proposition C.4.1 for a standard derivation for the dual form of a general φ-divergence distributionally robust performative risk.

Proof of Proposition 4.3.2. This follows Proposition 1 presented in Hu and Hong (2013) and the discussion paragraph right after the proposition. □

C.2.2 Proof of Proposition 4.3.3

Proof. We have the following decomposition of the excess risk bound of the PO,

$$
\begin{aligned}
&\mathrm{PR}_{\mathrm{true}}(\theta_{\mathrm{PO}}) - \min_{\theta\in\Theta} \mathrm{PR}_{\mathrm{true}}(\theta) \\
=\ &\mathrm{PR}_{\mathrm{true}}(\theta_{\mathrm{PO}}) - \mathrm{PR}_{\mathrm{true}}(\theta^\star_{\mathrm{PO}}) \\
=\ &\mathrm{PR}_{\mathrm{true}}(\theta_{\mathrm{PO}}) - \mathrm{PR}(\theta_{\mathrm{PO}}) + \underbrace{\mathrm{PR}(\theta_{\mathrm{PO}}) - \mathrm{PR}(\theta^\star_{\mathrm{PO}})}_{\leq 0 \text{ by definition of } \theta_{\mathrm{PO}}} + \mathrm{PR}(\theta^\star_{\mathrm{PO}}) - \mathrm{PR}_{\mathrm{true}}(\theta^\star_{\mathrm{PO}}) \\
\leq\ &2 \sup_{\theta\in\Theta} \left| \mathrm{PR}(\theta) - \mathrm{PR}_{\mathrm{true}}(\theta) \right| \\
=\ &2 \sup_{\theta\in\Theta} \left| \mathbb{E}_{Z\sim\mathcal{D}(\theta)}[\ell(Z;\theta)] - \mathbb{E}_{Z\sim\mathcal{D}_{\mathrm{true}}(\theta)}[\ell(Z;\theta)] \right| \\
\leq\ &2B \sup_{\theta\in\Theta} D_{\mathrm{TV}}(\mathcal{D}_{\mathrm{true}}(\theta) \| \mathcal{D}(\theta)) \\
\leq\ &\sqrt{2}B \sup_{\theta\in\Theta} \sqrt{D_{\mathrm{KL}}(\mathcal{D}_{\mathrm{true}}(\theta) \| \mathcal{D}(\theta))}.
\end{aligned}
$$

The last two inequalities are due to the asumption of bounded loss, $|\ell(z;\theta)| \leq B$, and Pinksker's inequality. □

C.2.3 Proof of Proposition 4.3.4

Proof. We have the following decomposition of the excess risk bound of the DRPO,

$$
\begin{aligned}
& \mathrm{PR}_{\mathrm{true}}(\theta_{\mathrm{DRPO}}) - \min_{\theta\in\Theta} \mathrm{PR}_{\mathrm{true}}(\theta) \\
= \; & \mathrm{PR}_{\mathrm{true}}(\theta_{\mathrm{DRPO}}) - \mathrm{PR}_{\mathrm{true}}(\theta^\star_{\mathrm{PO}}) \\
= \; & \underbrace{\mathrm{PR}_{\mathrm{true}}(\theta_{\mathrm{DRPO}}) - \mathrm{DRPR}(\theta_{\mathrm{DRPO}})}_{\leq 0 \text{ by generalization principle}} + \underbrace{\mathrm{DRPR}(\theta_{\mathrm{DRPO}}) - \mathrm{DRPR}(\theta^\star_{\mathrm{PO}})}_{\leq 0 \text{ by definition of } \theta_{\mathrm{DRPO}}} \\
& + \mathrm{DRPR}(\theta^\star_{\mathrm{PO}}) - \mathrm{PR}_{\mathrm{true}}(\theta^\star_{\mathrm{PO}}) \\
\leq \; & \mathrm{DRPR}(\theta^\star_{\mathrm{PO}}) - \mathrm{PR}_{\mathrm{true}}(\theta^\star_{\mathrm{PO}}) \\
= \; & \sqrt{\mathrm{Var}_{Z\sim\mathcal{D}(\theta^\star_{\mathrm{PO}})}[\ell(Z;\theta^\star_{\mathrm{PO}})]\rho} + o(\sqrt{\rho}).
\end{aligned}
$$

The last equality is due to the sensitivity property of KL divergence-based DRO (Duchi et al., 2021). □

C.2.4 Proof of Proposition 4.3.5

Proof. For any distribution P such that $D_{\mathrm{KL}}(P\|\mathcal{D}(\theta_{\mathrm{DRPO}})) \le \rho$, we have the following decomposition of the excess risk bound of the DRPO,

$$
\begin{aligned}
&\mathrm{PR}_{\mathrm{true}}(\theta_{\mathrm{DRPO}}) - \min_{\theta\in\Theta} \mathrm{PR}_{\mathrm{true}}(\theta) \\
=\ & \mathrm{PR}_{\mathrm{true}}(\theta_{\mathrm{DRPO}}) - \mathrm{PR}_{\mathrm{true}}(\theta^\star_{\mathrm{PO}}) \\
=\ & \mathrm{PR}_{\mathrm{true}}(\theta_{\mathrm{DRPO}}) - \mathbb{E}_{Z\sim P}[\ell(Z;\theta_{\mathrm{DRPO}})] + \underbrace{\mathbb{E}_{Z\sim P}[\ell(Z;\theta_{\mathrm{DRPO}})] - \mathrm{DRPR}(\theta_{\mathrm{DRPO}})}_{\le 0 \text{ by generalization principle}} \\
&+ \underbrace{\mathrm{DRPR}(\theta_{\mathrm{DRPO}}) - \mathrm{DRPR}(\theta^\star_{\mathrm{PO}})}_{\le 0 \text{ by definition of } \theta_{\mathrm{DRPO}}} + \mathrm{DRPR}(\theta^\star_{\mathrm{PO}}) - \mathrm{PR}_{\mathrm{true}}(\theta^\star_{\mathrm{PO}}) \\
\le\ & \mathrm{PR}_{\mathrm{true}}(\theta_{\mathrm{DRPO}}) - \mathbb{E}_{Z\sim P}[\ell(Z;\theta_{\mathrm{DRPO}})] + \mathrm{DRPR}(\theta^\star_{\mathrm{PO}}) - \mathrm{PR}_{\mathrm{true}}(\theta^\star_{\mathrm{PO}}) \\
=\ & \mathrm{PR}_{\mathrm{true}}(\theta_{\mathrm{DRPO}}) - \mathbb{E}_{Z\sim P}[\ell(Z;\theta_{\mathrm{DRPO}})] + \sqrt{\mathrm{Var}_{Z\sim\mathcal{D}(\theta^\star_{\mathrm{PO}})}[\ell(Z;\theta^\star_{\mathrm{PO}})]\rho} + o(\sqrt{\rho}) \\
=\ & \mathbb{E}_{Z\sim\mathcal{D}_{\mathrm{true}}(\theta_{\mathrm{DRPO}})}[\ell(Z;\theta_{\mathrm{DRPO}})] - \mathbb{E}_{Z\sim P}[\ell(Z;\theta_{\mathrm{DRPO}})] + \sqrt{\mathrm{Var}_{Z\sim\mathcal{D}(\theta^\star_{\mathrm{PO}})}[\ell(Z;\theta^\star_{\mathrm{PO}})]\rho} \\
&+ o(\sqrt{\rho}) \\
\le\ & B D_{\mathrm{TV}}(\mathcal{D}_{\mathrm{true}}(\theta_{\mathrm{DRPO}})\|P) + \sqrt{\mathrm{Var}_{Z\sim\mathcal{D}(\theta^\star_{\mathrm{PO}})}[\ell(Z;\theta^\star_{\mathrm{PO}})]\rho} + o(\sqrt{\rho}) \\
\le\ & \sqrt{2} B D_{\mathrm{KL}}(\mathcal{D}_{\mathrm{true}}(\theta_{\mathrm{DRPO}})\|P) + \sqrt{\mathrm{Var}_{Z\sim\mathcal{D}(\theta^\star_{\mathrm{PO}})}[\ell(Z;\theta^\star_{\mathrm{PO}})]\rho} + o(\sqrt{\rho}).
\end{aligned}
$$

By the arbitrariness of P in the divergence ball $D_{\mathrm{KL}}(P\|\mathcal{D}(\theta_{\mathrm{DRPO}})) \le \rho$, we complete the proof. □

C.3 Experimental Details and More Results

For the loss function $\ell(x, y; \theta)$, we adopt the cross-entropy loss with L_2-regularization, that is,

$$
\ell(x, y; \theta) = -y \log h_\theta(x) - (1 - y) \log(1 - h_\theta(x)) + \frac{\lambda}{2}\|\theta\|_2^2,
$$

where $h_\theta(x) = \left(1 + \exp\{-\theta^\top x\}\right)^{-1}$ and $\lambda = 0.001$. Given a model θ, we predict $\widehat{y} = \mathbf{1}\{h_\theta(x) \ge 0.5\}$.

C.3.1 Strategic Classification with Misspecified Cost Function in Section 4.5.1 (Continued)

The data preprocessing procedure for the credit dataset Kaggle (2012) follows the procedure documented in Perdomo et al. (2020). After that procedure, the base distribution $\mathcal{D}_{\text{true}}(\mathbf{0})$ has $n = 14878$ data points with equal probability mass. We treat the distribution map associated with $\mathcal{D}_{\text{true}}(\mathbf{0})$ as the underlying test distribution map which is unknown to us. We generate n IID samples from $\mathcal{D}_{\text{true}}(\mathbf{0})$ to get a training base distribution $\widehat{\mathcal{D}}(\mathbf{0})$, that is, $\widehat{\mathcal{D}}(\mathbf{0}) \sim \mathcal{D}_{\text{true}}(\mathbf{0})^{\otimes n}$, and then we have a training nominal distribution map induced by $\widehat{\mathcal{D}}(\mathbf{0})$. This training procedure is repeated for 10 trials, and the shaded region in each of the figures in this chapter represents the standard error of the mean calculated from the 10 trials for the corresponding curve.

The left of Figure C.1 shows performative balanced error rate, which refers to the balanced error rate (BER) on the test model-induced distribution, incurred by the PO and the DRPO's with various radius ρ's. Because the cross-entropy loss serves as a surrogate for classification error, we see patterns of these curves similar to those in the left of Figure IV.1: as ρ increases, the DRPO achieves more uniform performance across a wider range of ϵ_{true}. On the other hand, because the performative classification error is not exactly the criterion we minimize, different patterns are also observed: the DRPO with relatively large radius $\rho = 0.04$ outperforms the PO for all $\epsilon_{\text{true}} \in [0, 1]$. As indicated by the right of Figure C.1, by increasing ρ the DRPO constantly improves the relative worst-case performance in terms of performative BER for $\eta \in \{0.6, 0.8, 1.0\}$. Here the relative worst-case improvement in performative BER of θ to θ_{PO} is defined by $\frac{\max_{\epsilon_{\text{true}}} \text{BER}_{\text{true}}(\theta_{\text{PO}}) - \max_{\epsilon_{\text{true}}} \text{BER}_{\text{true}}(\theta)}{\max_{\epsilon_{\text{true}}} \text{BER}_{\text{true}}(\theta_{\text{PO}})} \times 100\%$.

For implementing the post-fitting calibration procedure in Section 4.4.4, we have to estimate the KL divergence between $\mathcal{D}_{\text{true}}(\theta_{\text{DRPO}}(\rho))$ and $\mathcal{D}(\theta_{\text{DRPO}}(\rho))$ for any fixed ρ. Here we index θ_{DRPO} by the radius ρ to emphasize its dependence on ρ. Recall

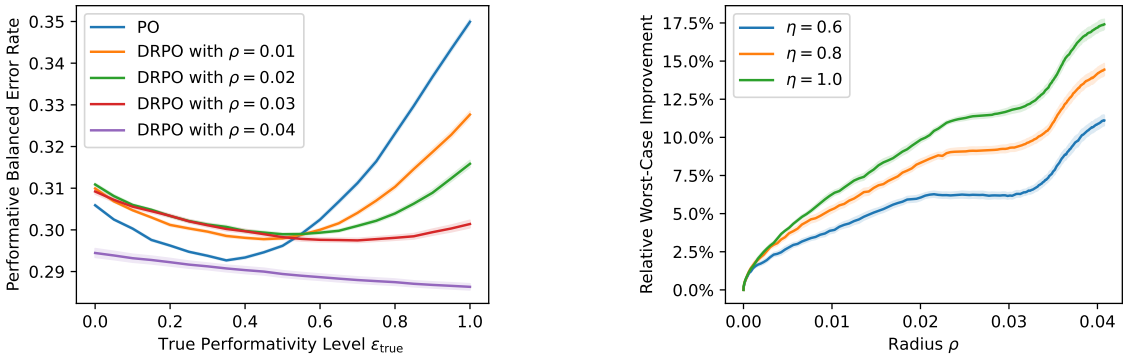

Figure C.1: Additional results of Experiment 4.5.1. Left: performative balanced error rate incurred by the PO and the DRPO's with various radius ρ's. Right: relative improvement in worst-case performative balanced error rate of the DRPO to the PO as the radius ρ increases, for different range of misspecification η's.

that

$$\begin{aligned} Z \sim \mathcal{D}(\theta) &\iff Z \overset{d}{=} Z_0 - 0.5A\theta \text{ and } Z \sim \mathcal{D}_{\text{true}}(\theta) \\ &\iff Z \overset{d}{=} Z_0 - \epsilon_{\text{true}} A\theta \text{ where } Z_0 \sim \mathcal{D}_{\text{true}}(\mathbf{0}). \end{aligned}$$

Here $A^\top = \left[B^\top \mid \mathbf{0}_d\right]$ (see Section 4.5.2 for the definition of B) and we have training base distribution $\widehat{D}(\mathbf{0})$ from $\mathcal{D}_{\text{true}}(\mathbf{0})$. We use an inexact parametric method to estimate the KL divergence. Assume $\mathcal{D}_{\text{true}}(\mathbf{0}) \sim \mathcal{N}(\mu, \Sigma)$, then

$$\begin{aligned} D(\mathcal{D}_{\text{true}} \| \mathcal{D}) &= \mathcal{D}(\mathcal{N}(\mu - 0.5A\theta_{\text{DRPO}}, \Sigma) \| \mathcal{N}(\mu - \epsilon_{\text{true}} A\theta_{\text{DRPO}}, \Sigma)) \\ &= 0.5(\epsilon_{\text{true}} - 0.5)^2 \theta_{\text{DRPO}}^\top A^\top \Sigma^{-1} A\theta_{\text{DRPO}}. \end{aligned}$$

Let $\widehat{\Sigma}$ be the sample covariance of $\widehat{D}(\mathbf{0})$, a plug-in method implies an estimate of the KL divergence, which is given by

$$\widehat{D}(\mathcal{D}_{\text{true}}(\theta_{\text{DRPO}}(\rho)) \| \mathcal{D}(\theta_{\text{DRPO}}(\rho))) = 0.5(\epsilon_{\text{true}} - 0.5)^2 \theta_{\text{DRPO}}^\top A^\top \widehat{\Sigma}^{-1} A\theta_{\text{DRPO}}.$$

C.3.2 Partially Identifiable Distribution Map in Section 4.5.2 (Continued)

The data generating procedure is the same as that in Experiment 4.5.1. The left of Figure C.2 shows performative balanced error rate (BER) incurred by the PO and the TPO's with various tilt α's. Similar to the left of Figure C.1, we observe that 1) as α increases, the TPO achieves more uniform performance across a wider range of ϵ_{true}; 2) the TPO with relatively large tilt $\alpha = 0.5$ outperforms the PO for all $\epsilon_{\text{true}} \in [-1, 1]$. The right of Figure C.2 shows relative improvement in worst-case performative balanced error rate of the TPO to the PO as the tilt α increases, for different range of misspecification η's. Without misspeicification, $\eta = 0$, the TPO has comparable performance to the PO in terms of performative BER. With moderate to large misspecification, $\eta \in \{2/3, 4/3, 2\}$, the TPO shows significant advantages over the PO in terms of relative improvement in worst-case performative BER.

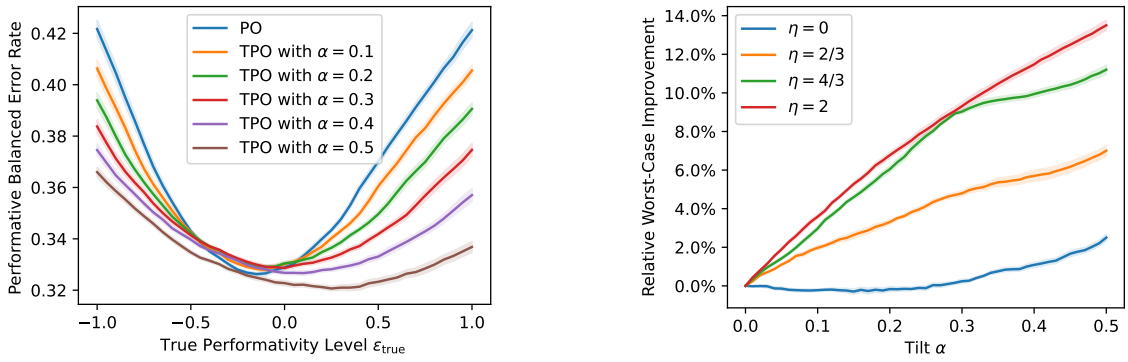

Figure C.2: Additional results of Experiment 4.5.2. Left: performative balanced error rate incurred by the PO and the TPO's with various tilt α's. Right: relative improvement in worst-case performative balanced error rate of the TPO to the PO as the tilt α increases, for different range of misspecification η's.

C.3.3 Fairness without Demographics in Section 4.5.3 (Continued)

We adopt the following data generating process similar to that in Peet-Pare et al. (2022). Let $X \sim \gamma\mathcal{N}(\mu_A, \Sigma_A) + (1-\gamma)\mathcal{N}(\mu_B, \Sigma_B)$. Let $\gamma = 0.8$ so that group A is the

majority group and group B is the minority group. Let $\mu_A = 1 \times \mathbf{1}_d$, $\mu_B = 0.8 \times \mathbf{1}_d$, and $\Sigma_A = \Sigma_B = 0.1 \times I_d$. If X comes from group A, then label $Y = 0$ if $X^\top \mathbf{1}_d \leq \mu_A^\top \mathbf{1}_d$. If X comes from group B, then label $Y = 0$ if $X^\top \mathbf{1}_d \leq \mu_B^\top \mathbf{1}_d$. The distribution map follows:

$$X_{1:\lfloor d/2 \rfloor} \leftarrow X_{1:\lfloor d/2 \rfloor} - \epsilon \theta_{1:\lfloor d/2 \rfloor}$$
$$X_{(\lfloor d/2 \rfloor + 1):d} \leftarrow X_{(\lfloor d/2 \rfloor + 1):d}$$

so that the first $\lfloor d/2 \rfloor$ features are strategic features, and ϵ controls the strength of performativity. We choose $d = 10$ and $\epsilon = 0.5$ to wrap up the setup. Finally, we assume knowledge of the true performativity and observe IID samples of size $n = 12500$ from the population base distribution. In short, we eliminate the effect of population distribution map misspecification, and instead concentrate on the effect of subpopulation distribution map shift.

Figure C.3 shows the performative accuracy, which refers to the accuracy on the model-induced distribution, of the population, the majority, and the minority incurred by the TPO, as the tilt α increases. Because the cross-entropy loss serves as a surrogate for classification error, we see patterns of the three curves similar to those in Figure IV.4. The PO (which is TPO with $\alpha = 0$) exhibits the highest performative accuracy at the population, but the greatest disparity between its performance for the majority and minority groups. As α increases, the TPO reduces the performance gap between the two groups at the expense of an decreased population performative accuracy. This suggests that the distributionally robust performative prediction framework has the potential to mitigate unfairness towards the minority group, even in the absence of demographic information.

For implementing the calibration set procedure in Section 4.4.4, we have to obtain access to a few samples with the group membership information. Here we allow

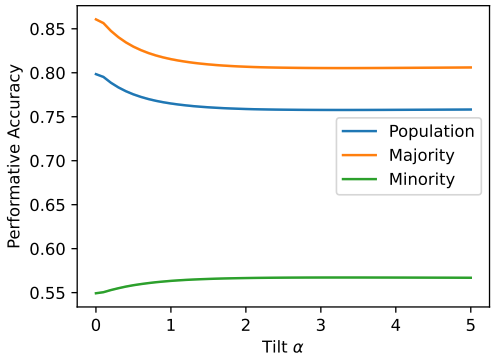

Figure C.3: Additional results of Experiment 4.5.3. Performative accuracy of the population, the majority, and the minority, as the tilt α increases.

ourselves to observe IID samples of size $n_{\text{cal}} = 125$ with the known group membership from the population base distribution.

C.4 Extension to General φ-Divergence

The KL divergence distributionally robust performative prediction framework can be extended to a general φ-divergence distributionally robust performative prediction framework. Recall that a φ-divergence is defined by

$$D_\varphi(Q\|P) = \int \varphi\left(\frac{dQ}{dP}\right) dP,$$

where $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ and $\varphi(1) = 0$. Here dQ/dP is the Radon–Nikodym derivative, and we implicitly require the probability measure Q to be absolutely continuous with respect to P. Note that if we choose $\varphi(t) = t \log t$, then we recover the framework presented in the main article. Now we keep φ as a generic function and we will instantiate some popular families of φ-divergence after the presentation of the general framework.

With the φ-divergence, we can define a family of distribution maps around the nominal distribution map. Specifically, the *uncertainty collection* around $\mathcal{D}$ with

critical radius ρ is defined as

$$\mathcal{U}(\mathcal{D}) = \{\widetilde{\mathcal{D}} : \Theta \to \mathcal{P}(\mathcal{Z}) \mid D_\varphi(\widetilde{\mathcal{D}}(\theta) \| \mathcal{D}(\theta)) \leq \rho, \forall \theta \in \Theta\}.$$

Definition C.4.1 (φ-divergence distributionally robust performative risk)**.** The φ-divergence distributionally robust performative risk with the uncertainty collection $\mathcal{U}(\mathcal{D})$ is defined as

$$\mathrm{DRPR}_\varphi(\theta) = \sup_{\widetilde{\mathcal{D}}:\widetilde{\mathcal{D}} \in \mathcal{U}(\mathcal{D})} \mathbb{E}_{Z\sim\widetilde{\mathcal{D}}(\theta)}[\ell(Z;\theta)]. \tag{C.4.1}$$

The evaluation of φ-divergence distributionally robust performative risk $\mathrm{DRPR}_\varphi(\theta)$ given in (C.4.1) involves an infinite dimensional maximization problem which is generally intractable. Fortunately, it is in fact equivalent to a minimization problem over two dual variables. This is given by the following strong duality result.

Proposition C.4.1 (Strong duality of DRPR_φ)**.** *For any $\theta \in \Theta$, we have*

$$\mathrm{DRPR}_\varphi(\theta) = \inf_{\mu\geq 0, \nu\in\mathbb{R}} \left\{ \mathbb{E}_{Z\sim\mathcal{D}(\theta)} \left[\mu\varphi^* \left(\frac{\ell(Z;\theta) - \nu}{\mu} \right) + \mu\rho + \nu \right] \right\}, \tag{C.4.2}$$

where $\varphi^*(s) = \sup_t\{st - \varphi(t)\}$ is the convex conjugate of φ.

Proof. We only have to show that for any fixed $\theta \in \Theta$, the dual form (C.4.2) holds. We introduce the likelihood ratio $L(Z) = d\widetilde{\mathcal{D}}(\theta)/d\mathcal{D}(\theta)$. By change of variable, we can rewrite the φ-divergence distributionally robust performative risk (C.4.1) as

$$\begin{aligned} &\sup_{\widetilde{\mathcal{D}}(\theta):D_\varphi(\widetilde{\mathcal{D}}(\theta)\|\mathcal{D}(\theta))\leq\rho} \mathbb{E}_{Z\sim\widetilde{\mathcal{D}}(\theta)}[\ell(Z;\theta)] \\ = &\sup_{L\geq 0} \left\{ \mathbb{E}_{Z\sim\mathcal{D}(\theta)}[L(Z)\ell(Z;\theta)] \mid \mathbb{E}_{Z\sim\mathcal{D}(\theta)}[\varphi(L(Z))] \leq \rho, \mathbb{E}_{Z\sim\mathcal{D}(\theta)}[L(Z)] = 1 \right\}, \end{aligned}$$

where the supremum takes over measurable functions. This gives us a constrained optimization problem. Let $\mu \geq 0$ be the Lagrange multiplier for $\mathbb{E}_{Z\sim\mathcal{D}(\theta)}[\varphi(L(Z))] \leq \rho$ and $\nu \in \mathbb{R}$ be the Lagrange multiplier for $\mathbb{E}_{Z\sim\mathcal{D}(\theta)}[L(Z)] = 1$. The corresponding

Lagrangian is

$$\mathcal{L}(L,\mu,\nu) = \mathbb{E}_{Z\sim\mathcal{D}(\theta)}[(\ell(Z;\theta)-\nu)L(Z) - \mu\varphi(L(Z))] + \mu\rho + \nu.$$

For regular φ-divergence, we have

$$\begin{aligned}
&\sup_{\widetilde{\mathcal{D}}(\theta):D_\varphi(\widetilde{\mathcal{D}}(\theta)\|\mathcal{D}(\theta))\le\rho} \mathbb{E}_{Z\sim\widetilde{\mathcal{D}}(\theta)}[\ell(Z;\theta)] \\
&= \inf_{\mu\ge 0,\nu\in\mathbb{R}} \sup_{L\ge 0} \mathcal{L}(L,\mu,\nu) \\
&= \inf_{\mu\ge 0,\nu\in\mathbb{R}} \sup_{L\ge 0} \left\{\mathbb{E}_{Z\sim\mathcal{D}(\theta)}[(\ell(Z;\theta)-\nu)L(Z) - \mu\varphi(L(Z))] + \mu\rho + \nu\right\} \\
&= \inf_{\mu\ge 0,\nu\in\mathbb{R}} \sup_{L\ge 0} \left[\mathbb{E}_{Z\sim\mathcal{D}(\theta)}\left[\mu\left\{\frac{\ell(Z;\theta)-\nu}{\mu}L(Z) - \varphi(L(Z))\right\} + \mu\rho + \nu\right]\right\} \\
&= \inf_{\mu\ge 0,\nu\in\mathbb{R}} \left\{\mathbb{E}_{Z\sim\mathcal{D}(\theta)}\left[\mu\sup_{t\ge 0}\left\{\frac{\ell(Z;\theta)-\nu}{\mu}t - \varphi(t)\right\} + \mu\rho + \nu\right]\right\} \\
&= \inf_{\mu\ge 0,\nu\in\mathbb{R}} \left\{\mathbb{E}_{Z\sim\mathcal{D}(\theta)}\left[\mu\varphi^*\left(\frac{\ell(Z;\theta)-\nu}{\mu}\right) + \mu\rho + \nu\right]\right\}.
\end{aligned}$$

Here the last equality holds according to the definition of the convex conjugate $\varphi^*(\cdot)$.

□

To develop an algorithm for finding the DRPO, we introduce the (μ,ν)-*augmented performative risk*.

Definition C.4.2 ((μ,ν)-augmented performative risk)**.** The (μ,ν)-augmented performative risk is defined by

$$\mathrm{AugPR}_\varphi(\theta,\mu,\nu) = \mathbb{E}_{Z\sim\mathcal{D}(\theta)}\left[\mu\varphi^*\left(\frac{\ell(Z;\theta)-\nu}{\mu}\right) + \mu\rho + \nu\right]. \tag{C.4.3}$$

From (C.4.2) and (C.4.3), we see that the minimization problem of $\mathrm{DRPR}_\varphi(\theta)$ over θ is equivalent to the minimization problem of $\mathrm{AugPR}_\varphi(\theta,\mu,\nu)$ jointly over (θ,μ,ν), which is itself a performative risk minimization problem. We summarize this procedure in Algorithm 7.

Algorithm 7 Augmented Performative Risk Minimization

1: **Input:** nominal distribution map $\mathcal{D}(\theta)$
2: Update $(\theta, \mu, \nu) \leftarrow \arg\min_{\theta\in\Theta,\mu\geq 0,\nu\in\mathbb{R}} \left\{\mathbb{E}_{Z\sim\mathcal{D}(\theta)}\left[\mu\varphi^*\left(\frac{\ell(Z;\theta)-\nu}{\mu}\right)+\mu\rho+\nu\right]\right\}$
3: **Return:** θ

For special choice of φ, the strong duality result (C.4.1) can be reduced to involving only a single dual variable. Now we instantiate φ-divergence as the Cressie-Read family:

$$\phi_k(t) = \frac{t^k - kt + k - 1}{k(k-1)},$$

for $k > 1$. Like KL divergence, a distributionally robust performative prediction problem induced by a divergence from the Cressie-Read family has a dual reformualtion with single dual variable:

$$\mathrm{DRPR}_{\phi_k}(\theta) = \inf_{\mu\geq 0}\left\{(1+\rho k(k-1))^{1/k}\mathbb{E}_{Z\sim\mathcal{D}(\theta)}\left[\max\{\ell(Z;\theta)-\mu,0\}^{k_*}\right]^{1/k_*}+\mu\right\}, \tag{C.4.4}$$

where k_* is the conjugate number of k such that $1/k + 1/k_* = 1$. Therefore, an algorithm parallel to Algirthm 2 can be developed in a similar fashion based on the single-variable dual form (C.4.4).

As a final remark, all of the theoretical results regarding excess risk bounds (see Proposition 4.3.4 and Proposition 4.3.5) are still valid for the general φ-divergence distributionally robust performative prediction.

On the other hand, it is possible to extend Proposition 4.3.3 to general φ-divergence, and the result statement needs a slight modification. By generalized Pinsker's inequality, there is an increasing function $F : [0,2] \to \mathbb{R}_+$ such that $D_\varphi(P\|Q) \geq F(\mathrm{TV}(P\|Q))$. Then for general φ-divergence, Proposition 4.3.3 can be modified to

$$\mathcal{E}(\theta_{\mathrm{PO}}) \leq 2B \sup_{\theta\in\Theta} F^{-1}\left(D_\varphi(\mathcal{D}_{\mathrm{true}}(\theta)\|\mathcal{D}(\theta))\right)$$

or

$$\mathcal{E}(\theta_{\mathrm{PO}}) \leq 2BF^{-1}\left(\sup_{\theta\in\Theta} D_{\varphi}(\mathcal{D}_{\mathrm{true}}(\theta)\|\mathcal{D}(\theta))\right)$$

due to monotonicity of $F(\cdot)$. For a concrete example, $F(v) = v^2\mathbf{1}\{v < 1\} + \frac{v}{2-v}\mathbf{1}\{v \geq 1\}$ when $\varphi(t) = (t-1)^2$.

C.5 Discussion of Algorithmic Convergence

The convergence guarantees of the proposed algorithms can be established on a case-by-case basis, depending on the specific "inner" algorithm or solver employed for solving performative risk minimization. Existing algorithmic convergence results for model-based PO solvers (Miller et al., 2021; Izzo et al., 2021) can be taped into our algorithms naturally. Moreover, empirical results in Section 4.5 validate the effectiveness of our proposed algorithms.

BIBLIOGRAPHY

BIBLIOGRAPHY

Achiam, J., et al. (2023), Gpt-4 technical report, *arXiv preprint arXiv:2303.08774*.

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

Agarwal, A., M. Dudík, and Z. S. Wu (2019), Fair regression: Quantitative definitions and reduction-based algorithms, in *International Conference on Machine Learning*, pp. 120–129, PMLR.

Ahmadi-Javid, A. (2012), Entropic value-at-risk: A new coherent risk measure, *Journal of Optimization Theory and Applications*, *155*, 1105–1123.

Allison, P. D. (2001), *Missing data*, Sage publications.

Amodei, D., C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané (2016), Concrete Problems in AI Safety, *arXiv:1606.06565 [cs]*.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner (2022), Machine bias, in *Ethics of data and analytics*, pp. 254–264, Auerbach Publications.

Awasthi, P., M. Kleindessner, and J. Morgenstern (2020), Equalized odds postprocessing under imperfect group information, in *International conference on artificial intelligence and statistics*, pp. 1770–1780, PMLR.

Barocas, S., M. Hardt, and A. Narayanan (2023), *Fairness and machine learning: Limitations and opportunities*, MIT Press.

Bayraksan, G., and D. K. Love (2015), Data-driven stochastic programming using phi-divergences, in *The operations research revolution*, pp. 1–19, INFORMS.

Ben-Tal, A., D. den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen (2012), Robust Solutions of Optimization Problems Affected by Uncertain Probabilities, *Management Science*, *59*(2), 341–357, doi:10.1287/mnsc.1120.1641.

Bengio, Y., et al. (2023), Managing ai risks in an era of rapid progress, *arXiv preprint arXiv:2310.17688*.

Blanchet, J., and Y. Kang (2021), Sample out-of-sample inference based on wasserstein distance, *Operations Research*, *69*(3), 985–1013.

Blanchet, J., and K. Murthy (2019), Quantifying distributional model risk via optimal transport, *Mathematics of Operations Research*, *44*(2), 565–600.

Blanchet, J., Y. Kang, and K. Murthy (2019), Robust wasserstein profile inference and applications to machine learning, *Journal of Applied Probability*, *56*(3), 830–857.

Brown, G., S. Hod, and I. Kalemaj (2022), Performative prediction in a stateful world, in *International Conference on Artificial Intelligence and Statistics*, pp. 6045–6061, PMLR.

Celis, L. E., L. Huang, V. Keswani, and N. K. Vishnoi (2021), Fair classification with noisy protected attributes: A framework with provable guarantees, in *International Conference on Machine Learning*, pp. 1349–1361, PMLR.

Chen, J., N. Kallus, X. Mao, G. Svacha, and M. Udell (2019), Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved, *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, pp. 339–348, doi:10.1145/3287560.3287594.

Chen, R., and I. C. Paschalidis (2018), A robust learning approach for regression models based on distributionally robust optimization, *Journal of Machine Learning Research*, *19*(13).

Chen, X., M. Sim, and P. Sun (2007), A Robust Optimization Perspective on Stochastic Programming, *Operations Research*, *55*, 1058–1071, doi:10.1287/opre.1070.0441.

Cotter, A., M. Gupta, H. Jiang, N. Srebro, K. Sridharan, S. Wang, B. Woodworth, and S. You (2019a), Training well-generalizing classifiers for fairness metrics and other data-dependent constraints, in *International Conference on Machine Learning*, pp. 1397–1405, PMLR.

Cotter, A., H. Jiang, M. Gupta, S. Wang, T. Narayan, S. You, and K. Sridharan (2019b), Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals, *Journal of Machine Learning Research*, *20*(172), 1–59.

Cotter, A., H. Jiang, and K. Sridharan (2019c), Two-Player Games for Efficient Non-Convex Constrained Optimization, in *Algorithmic Learning Theory*, pp. 300–332, PMLR.

D'Amour, A., et al. (2022), Underspecification presents challenges for credibility in modern machine learning, *Journal of Machine Learning Research*, *23*(226), 1–61.

Davenport, M. A., R. G. Baraniuk, and C. D. Scott (2010), Tuning support vector machines for minimax and neyman-pearson classification, *IEEE transactions on pattern analysis and machine intelligence*, *32*(10), 1888–1898.

Davenport, T., and R. Kalakota (2019), The potential for artificial intelligence in healthcare, *Future healthcare journal*, *6*(2), 94.

Delage, E., and Y. Ye (2010), Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems, *Operations Research*, *58*, 595–612, doi:10.1287/opre.1090.0741.

Ding, F., M. Hardt, J. Miller, and L. Schmidt (2021), Retiring Adult: New Datasets for Fair Machine Learning, in *Advances in Neural Information Processing Systems*.

Dong, R., H. Zhang, and L. Ratliff (2023), Approximate regions of attraction in learning with decision-dependent distributions, in *International Conference on Artificial Intelligence and Statistics*, pp. 11,172–11,184, PMLR.

Drusvyatskiy, D., and L. Xiao (2023), Stochastic optimization with decision-dependent distributions, *Mathematics of Operations Research*, *48*(2), 954–998.

Dua, D., and C. Graff (2017), UCI machine learning repository.

Duchi, J., and H. Namkoong (2019), Variance-based regularization with convex objectives, *Journal of Machine Learning Research*, *20*(68), 1–55.

Duchi, J. C., M. I. Jordan, and M. J. Wainwright (2013), Local privacy and statistical minimax rates, in *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438, IEEE.

Duchi, J. C., P. W. Glynn, and H. Namkoong (2021), Statistics of robust optimization: A generalized empirical likelihood approach, *Mathematics of Operations Research*.

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

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006), Calibrating noise to sensitivity in private data analysis, in *Theory of cryptography conference*, pp. 265–284, Springer.

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

Eban, E., M. Schain, A. Mackey, A. Gordon, R. Rifkin, and G. Elidan (2017), Scalable learning of non-decomposable objectives, in *Artificial intelligence and statistics*, pp. 832–840, PMLR.

Fernholz, L. T. (2012), *Von Mises calculus for statistical functionals*, vol. 19, Springer Science & Business Media.

Gabriel, I. (2020), Artificial intelligence, values, and alignment, *Minds and machines*, *30*(3), 411–437.

Gao, R., and A. Kleywegt (2023), Distributionally robust stochastic optimization with wasserstein distance, *Mathematics of Operations Research*, *48*(2), 603–655.

Gao, R., X. Chen, and A. J. Kleywegt (2022), Wasserstein distributionally robust optimization and variation regularization, *Operations Research*.

Gianfrancesco, M. A., S. Tamang, J. Yazdany, and G. Schmajuk (2018), Potential biases in machine learning algorithms using electronic health record data, *JAMA internal medicine*, *178*(11), 1544–1547.

Goh, G., A. Cotter, M. Gupta, and M. P. Friedlander (2016), Satisfying Real-world Goals with Dataset Constraints, *Advances in Neural Information Processing Systems*, *29*.

Goh, J., and M. Sim (2010), Distributionally Robust Optimization and Its Tractable Approximations, *Operations Research*, *58*(4-part-1), 902–917, doi:10.1287/opre.1090.0795.

Gonzalez, M. F., J. F. Capman, F. L. Oswald, E. R. Theys, and D. L. Tomczak (2019), "where's the io?" artificial intelligence and machine learning in talent management systems, *Personnel Assessment and Decisions*, *5*(3), 5.

Goodfellow, I. J., J. Shlens, and C. Szegedy (2014), Explaining and Harnessing Adversarial Examples.

Gotoh, J.-y., M. J. Kim, and A. E. Lim (2018), Robust empirical optimization is almost the same as mean–variance optimization, *Operations research letters*, *46*(4), 448–452.

Gudivada, V., A. Apon, and J. Ding (2017), Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations, *International Journal on Advances in Software*, *10*(1), 1–20.

Gupta, V. (2019), Near-optimal bayesian ambiguity sets for distributionally robust optimization, *Management Science*, *65*(9), 4242–4260.

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

Hashimoto, T., M. Srivastava, H. Namkoong, and P. Liang (2018), Fairness without demographics in repeated loss minimization, in *International Conference on Machine Learning*, pp. 1929–1938, PMLR.

Hendrycks, D., N. Carlini, J. Schulman, and J. Steinhardt (2021), Unsolved Problems in ML Safety, *arXiv:2109.13916 [cs]*.

Hu, Z., and L. J. Hong (2013), Kullback-leibler divergence constrained distributionally robust optimization, *Available at Optimization Online*, *1*(2), 9.

Hüllermeier, E., and W. Waegeman (2021), Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods, *Machine learning*, *110*(3), 457–506.

Izzo, Z., L. Ying, and J. Zou (2021), How to learn when data reacts to your model: performative gradient descent, in *International Conference on Machine Learning*, pp. 4641–4650, PMLR.

Izzo, Z., J. Zou, and L. Ying (2022), How to learn when data gradually reacts to your model, in *International Conference on Artificial Intelligence and Statistics*, pp. 3998–4035, PMLR.

Jiang, R., and Y. Guan (2016), Data-driven chance constrained stochastic program, *Mathematical Programming*, *158*(1), 291–327.

Jumper, J., et al. (2021), Highly accurate protein structure prediction with alphafold, *Nature*, *596*(7873), 583–589.

Kaggle (2012), Give me some credit, `https://www.kaggle.com/c/GiveMeSomeCredit/data`.

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

Kairouz, P., S. Oh, and P. Viswanath (2016), Extremal mechanisms for local differential privacy, *The Journal of Machine Learning Research*, *17*(1), 492–542.

Kallus, N., X. Mao, and A. Zhou (2022), Assessing algorithmic fairness with unobserved protected class using data combination, *Management Science*, *68*(3), 1959–1981.

Keener, R. W. (2010), *Theoretical statistics: Topics for a core course*, Springer Science & Business Media.

Kleywegt, A. J., A. Shapiro, and T. Homem-de Mello (2002), The sample average approximation method for stochastic discrete optimization, *SIAM Journal on optimization*, *12*(2), 479–502.

Konstantinov, N., and C. H. Lampert (2022), On the impossibility of fairness-aware learning from corrupted data, in *Algorithmic Fairness through the Lens of Causality and Robustness workshop*, pp. 59–83, PMLR.

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

Lahoti, P., A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, and E. Chi (2020), Fairness without demographics through adversarially reweighted learning, *Advances in neural information processing systems*, *33*, 728–740.

Lam, H. (2019), Recovering Best Statistical Guarantees via the Empirical Divergence-Based Distributionally Robust Optimization, *Operations Research*, *67*(4), 1090–1105, doi:10.1287/opre.2018.1786.

Lam, H., and E. Zhou (2015), Quantifying uncertainty in sample average approximation, in *2015 Winter Simulation Conference (WSC)*, pp. 3846–3857, doi:10.1109/WSC.2015.7408541.

Lamy, A., Z. Zhong, A. K. Menon, and N. Verma (2019), Noise-tolerant fair classification, *Advances in neural information processing systems*, *32*.

Lee, J., and M. Raginsky (2018), Minimax statistical learning with wasserstein distances, *Advances in Neural Information Processing Systems*, *31*.

Li, Q., and H.-T. Wai (2022), State dependent performative prediction with stochastic approximation, in *International Conference on Artificial Intelligence and Statistics*, pp. 3164–3186, PMLR.

Li, T., A. Beirami, M. Sanjabi, and V. Smith (2020), Tilted Empirical Risk Minimization, in *International Conference on Learning Representations*.

Li, T., A. Beirami, M. Sanjabi, and V. Smith (2023), On tilted losses in machine learning: Theory and applications, *Journal of Machine Learning Research*, *24*(142), 1–79.

Lin, L., and T. Zrnic (2023), Plug-in performative optimization, *arXiv preprint arXiv:2305.18728*.

Luedtke, J., and S. Ahmed (2008), A Sample Approximation Approach for Optimization with Probabilistic Constraints, *SIAM Journal on Optimization*, *19*(2), 674–699, doi:10.1137/070702928.

Mendler-Dünner, C., J. Perdomo, T. Zrnic, and M. Hardt (2020), Stochastic optimization for performative prediction, *Advances in Neural Information Processing Systems*, *33*, 4929–4939.

Milani Fard, M., Q. Cormier, K. Canini, and M. Gupta (2016), Launch and Iterate: Reducing Prediction Churn, in *Advances in Neural Information Processing Systems*, vol. 29, Curran Associates, Inc.

Miller, J. P., J. C. Perdomo, and T. Zrnic (2021), Outside the echo chamber: Optimizing the performative risk, in *International Conference on Machine Learning*, pp. 7710–7720, PMLR.

Mitchell, S., E. Potash, S. Barocas, A. D'Amour, and K. Lum (2021), Algorithmic fairness: Choices, assumptions, and definitions, *Annual review of statistics and its application*, *8*, 141–163.

Mohajerin Esfahani, P., and D. Kuhn (2018), Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations, *Mathematical Programming*, *171*(1-2), 115–166.

Mohseni, S., H. Wang, C. Xiao, Z. Yu, Z. Wang, and J. Yadawa (2022), Taxonomy of machine learning safety: A survey and primer, *ACM Computing Surveys*, *55*(8), 1–38.

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

Namkoong, H., and J. C. Duchi (2016), Stochastic Gradient Methods for Distributionally Robust Optimization with F-divergences, in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 2216–2224, Curran Associates Inc., USA.

Natarajan, N., I. S. Dhillon, P. K. Ravikumar, and A. Tewari (2013), Learning with noisy labels, *Advances in neural information processing systems*, *26*.

Owen, A. (1990), Empirical Likelihood Ratio Confidence Regions, *The Annals of Statistics*, *18*(1), 90–120, doi:10.1214/aos/1176347494.

Peet-Pare, L., N. Hegde, and A. Fyshe (2022), Long term fairness for minority groups via performative distributionally robust optimization, *arXiv preprint arXiv:2207.05777*.

Perdomo, J., T. Zrnic, C. Mendler-Dünner, and M. Hardt (2020), Performative prediction, in *International Conference on Machine Learning*, pp. 7599–7609, PMLR.

Ray, M., L. J. Ratliff, D. Drusvyatskiy, and M. Fazel (2022), Decision-dependent risk minimization in geometrically decaying dynamic environments, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 8081–8088.

Rubinstein, R. Y., and A. Shapiro (1993), *Discrete event systems: Sensitivity analysis and stochastic optimization by the score function method*, vol. 13, Wiley.

Shafieezadeh Abadeh, S., P. M. Mohajerin Esfahani, and D. Kuhn (2015), Distributionally robust logistic regression, *Advances in neural information processing systems*, *28*.

Shafieezadeh-Abadeh, S., D. Kuhn, and P. M. Esfahani (2019), Regularization via mass transportation, *Journal of Machine Learning Research*, *20*(103), 1–68.

Silver, D., et al. (2016), Mastering the game of go with deep neural networks and tree search, *nature*, *529*(7587), 484–489.

Sinha, A., H. Namkoong, and J. Duchi (2017), Certifying Some Distributional Robustness with Principled Adversarial Training, *arXiv:1710.10571 [cs, stat]*.

Wagner, C., M. Strohmaier, A. Olteanu, E. Kıcıman, N. Contractor, and T. Eliassi-Rad (2021), Measuring algorithmically infused societies, *Nature*, *595*(7866), 197–204.

Wang, S., W. Guo, H. Narasimhan, A. Cotter, M. Gupta, and M. Jordan (2020), Robust Optimization for Fairness with Noisy Protected Groups, *Advances in Neural Information Processing Systems*, *33*, 5190–5203.

Wang, W., and S. Ahmed (2008), Sample average approximation of expected value constrained stochastic programs, *Operations Research Letters*, *36*(5), 515–519, doi: 10.1016/j.orl.2008.05.003.

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

Woodworth, B., S. Gunasekar, M. I. Ohannessian, and N. Srebro (2017), Learning Non-Discriminatory Predictors, in *Conference on Learning Theory*, pp. 1920–1953, PMLR.

Yurochkin, M., and Y. Sun (2020), SenSeI: Sensitive Set Invariance for Enforcing Individual Fairness, in *International Conference on Learning Representations*.

Zafar, M. B., I. Valera, M. Gomez Rodriguez, and K. P. Gummadi (2017), Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment, in *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pp. 1171–1180, International World Wide Web Conferences Steering Committee, Perth, Australia, doi:10.1145/3038912.3052660.

Zafar, M. B., I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi (2019), Fairness Constraints: A Flexible Approach for Fair Classification, *Journal of Machine Learning Research*, *20*(75), 1–42.