Beyond Static Classification: Long-term Fairness for Minority Groups via Performative Prediction and Distributionally Robust Optimization

by

Garnet Liam Peet-Paré

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Statistical Machine Learning

Department of Mathematical and Statistical Sciences
University of Alberta

Abstract

In recent years machine learning (ML) models have begun to be deployed at enormous scales, but too often without adequate concern for whether or not an ML model will make fair decisions. Fairness in ML is a burgeoning research area, but work to define formal fairness criteria has some serious limitations. This thesis aims to combine and explore two recent areas of research in ML – distributionally robust optimization (DRO) and performative prediction – in an attempt to resolve some of these limitations. Performative prediction is a recent framework developed to understand the effects of when deploying model influences the distribution on which it is making predictions, an important concern for fairness. Research on performative prediction has thus far only examined risk minimization, however, which has the potential to result in discriminatory models when working with heterogeneous data composed of majority and minority subgroups. We examine performative prediction with a distributionally robust objective and prove an analogous convergence result for what we call repeated distributionally robust optimization (RDRO). We then verify our results empirically and develop experiments to demonstrate the impact of using RDRO on learning fair ML models.

Preface

This thesis is an original work by Garnet Liam Peet-Paré. No part of this thesis has been previously published.

To my parents and my sister, who told me to follow my curiosity, encouraged me to explore, and, most importantly, taught me compassion

Acknowledgements

As with most endeavors in my life, my experience in my master's degree has been shaped and defined primarily by the people I have met during my time in Edmonton. I chose to pursue graduate studies at the U of A with the intention of immersing myself in the ideas of machine learning, and while I certainly did that, the most rewarding aspect of my tenure here has not been my academic journey, but rather the friends that I made along the way.

I am lucky to have had three kind and patient supervisors in Alona Fyshe, Nidhi Hegde, and Adam Kashlak. While they were often not well equipped to provide me with academic guidance as I moved into research areas they were unfamiliar with, they were always supportive and encouraged me to pursue research that was meaningful to me. I am also grateful to Martha White and Csaba Svepesvari for their friendship and advice on various research topics I pursued.

I am enormously indebted to my primary research collaborators during my master's, Kirby Banman and Alan Chan. It was a joy to exchange ideas with them, admire their brilliance, and inevitably have our research discussions descend into fits of laughter.

To my closest friends here in Edmonton, who kept me sane during a lonely year of Covid, Tom Pinder, Katie Burak, Alexis Arrigoni, Bedir Tapkan, Nikoo Aghaei, and Parnian Mehinrad. Your friendship means more to me than you can imagine.

Some I wish would have met earlier in my master's, but even in the short time I've known them, I hope to have as lifelong friends: David Tao, Brad Burega, Alex Ayoub, Rohan Nuttall, Vlad Tkachuk, and Gabor Mihucz.

There are so many others who made my time in Edmonton one that I will remember fondly. There are too many to list, but some of them, in no particular order, are Rohan Saha, Cassidy Pirlot, Paul Mclaughlin, Liam Welsh, Mohsen Sattarifard, Kiarash Aghakasiri, Farnaz Kokhankhaki, Dhawal Gupta, Andy Patterson, Shibhansh Dohare, Jincheng Mei, Erfan Miahi, Shivam Garg, Connor Stephens, Shalaleh Rismani, Roshan Shariff, and Johannes Kirschner.

Table of Contents

List of Tables

List of Figures

"The true measure of any society can be found in how it treats its most vulnerable members."

-Mahatma Gandhi (possibly apocryphal)

Chapter 1

Introduction

In the past two decades, machine learning (ML) has moved from the confines of research institutes and university laboratories into the mainstream. Artificial intelligence (AI) is no longer merely the subject of science fiction films, but is now a crucial piece of technology that underpins some of the world's largest corporations and has become integral to people's everyday existence. At the time of writing, 10 of the top 11 world's most valuable publicly traded corporations had business models in which machine learning was a core component (Apple, Microsoft, Amazon, Alphabet, Meta, Tesla, Tencent, Alibaba Group, TSMC, Nvidia). Far from being fringe, AI has been woven into the very fabric of our society and economy and this trend will likely only accelerate in the coming years.

This move from the laboratory to the real world was spurred by the incredible successes of ML on practical problems. Deep learning, in particular, has demonstrated its spectacular ability to solve problems that were once thought too challenging for machines. While machine learning and deep learning have seen stunning successes and show the promise of helping to solve extremely important problems, the rapid rise of AI has far outpaced concern for questions of fairness, ethics, and governance of machine learning systems and algorithms. AI has already reshaped our world in remarkable ways and has made a select group of people incredibly wealthy, but the optimism and promise of AI has slowly begun to be replaced with pessimism and fear

of what unconstrained AI will do to our society. Many ML based companies operate in a regulatory Wild West and have thus far not demonstrated a willingness or ability to act responsibly, choosing instead to maximize profit at potentially great cost to society. Meta (formerly Facebook) has been the poster boy for this reckless behaviour based on profit maximization, and internal files leaked by former employee Frances Haugen to the Wall Street Journal revealed that Meta was aware of their platforms contributing to spreading misinformation prior to elections, fuelling hate connected to genocide in Myanmar, contributing to mental health issues in young women, and spreading misinformation regarding Covid vaccines that resulted in thousands of preventable deaths. The spread of information on Meta's platforms, and the extreme societal harm associated with it, is directly connected to their deployment of machine learning algorithms that recommend content to users [1].

Disturbingly, this lack of concern for the consequences of deploying AI algorithms at scale exists not just at the regulatory level, but also among many ML researchers and practitioners. The notion of making the consideration of potential harmful effects of AI models a central piece of the research process has yet to be accepted by the mainstream community. The relative lack of interest in questions of fairness and ethics among AI researchers is exacerbated by the inability of researchers in adjacent fields to effectively and adequately understand AI algorithms due to their technical complexity.

Significant interest in research into questions of fairness and ethics in machine learning only began several years ago and the field is still in its infancy. This research currently lags far behind the reality of how AI models are being deployed, with the result that we lack the tools to understand and predict the consequences of deploying AI algorithms. We all interact with AI on a daily basis and many of these interactions have fairness and ethical concerns associated with them, yet as a field we still have little understanding of these issues.

Deployment of ML models comes with ethical risks, but automated decision making

also offers an opportunity to make progress on ethical and fairness issues, as ML algorithms have the potential to uncover complex relationships that are essential to decision making that humans would otherwise overlook [2]. Machine learning also offers us the opportunity to provide transparency and consistency in decision making that is otherwise impossible to achieve with human decision makers. If this promise is to be fulfilled however, significant investment in fairness research is required.

The incentives within the field and market have thus far driven investment into areas that are financially profitable, which has not necessarily aligned with what benefits humanity. Algorithms are being deployed at enormous scales with shockingly little concern for the ramifications. According to its own data, Meta, for instance, reaches 3.6 billion people each month. With this many people affected in intimate ways by AI algorithms, it is essential that we develop a better understanding of ethics and governance of AI.

Ethics and governance of AI algorithms is a diverse and complex field, but one small part in which some progress has been made is in the area of fairness. Fairness in this context means ensuring that an AI model does not discriminate against particular subsets of people in its decision making. The subsets considered are usually legally protected classes such as race or gender, but fairness does not necessarily have to be restricted to characteristics that receive protection from discrimination under the law. At first glance it may seem like training a fair ML model is a simple task, but it turns out there are many subtle questions that complicate the matter and complex interactions between mathematical and ethical principles.

This thesis will present an introduction to research on fairness in machine learning, discuss two recent research areas with implications for fairness in ML, and extend the results of these areas both through theoretical and empirical studies. The goal of this work is to deepen our understanding of fairness in machine learning and to present a compelling alternative method for ensuring fairness in ML models.

More specifically, this thesis will proceed as follows. First, we will provide back-

ground on established definitions of what it means for an algorithm to be considered fair. As we explain, although this is fundamentally a contested notion, the field has largely coalesced around several formal fairness criteria, which are generally conceptually and mathematically straightforward, but have some serious limitations. We will then discuss two potential objectives for optimization, *empirical risk minimization* and *distributionally robust optimization*, and how they have implications for fairness in machine learning. Next, we cover an emerging area of research known as *performative prediction*, which attempts to provide a framework to capture the complexities of fairness dynamics over time, where feedback loops that occur as a result of a model affecting the population on which it makes predictions are a concern. After covering this background we discuss some related work that has inspired the research in this thesis by attempting to answer similar questions to those which we address here.

In Chapters 3 and 4 we introduce our novel theoretical and empirical contributions. We extend performative prediction to consider the distributionally robust objective by modifying the appropriate definitions and theorems. Replacing risk minimization with distributionally robust optimization results in additional technical complexity which we discuss. In Chapter 4, we investigate performative prediction with a distributionally robust objective through a series of experiments. We first reproduce an experiment from Perdomo *et al.* [3] to demonstrate the convergence properties of repeated risk minimization and repeated distributionally robust optimization. We then conduct a series of experiments on simple synthetic datasets to develop intuition for how risk minimization and distributionally robust optimization differ in the performative prediction context. Finally, we design an experiment to explore how distributionally robust optimization intersects with fairness concerns in machine learning and demonstrate that it has the potential for helping to ensure we train models that are non-discriminatory.

Chapter 2

Background

Fairness in ML is a diverse and multi-faceted topic that can mean many things. Kate Crawford in her Neurips 2017 keynote titled *The Trouble With Bias* gave a rough taxonomy of fairness concerns as falling into one of two categories: allocative harms or representative harms. Allocative harms refers to algorithms that make decisions about allocating resources whereas representational harms refers to algorithms that reinforce or amplify subordination of a group. These categories do not necessarily capture all scenarios in which fairness concerns arise, but they at least represent an attempt at creating a conceptual framework through which to view fairness concerns in ML. For instance, diffuse societal harms that can result from recommender systems deployed at scale are not easily captured by these categories.

2.1 Formal Fairness Criteria

Research into fairness in ML has thus far largely centred around questions of discrimination against certain protected demographic groups in classification tasks. These protected groups or classes are often based on characteristics that receive legal protection from discrimination, such as gender or race. This type of research tends to apply more to allocative harms, as classification algorithms are often making decisions about allocating resources among individuals, although this is not always the case depending on the application (*e.g.* facial recognition). The discussion of fairness

in this thesis will apply to questions of allocative harms, although there is potential for the techniques we discuss to be useful in combating discrimination in a wider variety of contexts.

There is no single, universal definition of what it means for an algorithm, even a classification algorithm, to be fair. Fairness is fundamentally a philosophical and political concept and, as such, depends heavily on context. That being said, there are a variety of mathematical definitions for fairness that have been proposed. While there are literally dozens of proposed definitions, many of these have been shown to be equivalent, or relaxations of each other.

The Fairness and Machine Learning textbook, authored by pioneers of ML fairness research [2], states that there are three central definitions of what constitutes a fair classifier. These definitions are known, respectively, as *independence*, *separation*, and *sufficiency*, and most other definitions for fair classifiers can be shown to be equivalent to, or a relaxed form of, one of these three definitions. These formal non-discrimination criteria have been developed by a number of researchers, often independently [4–11]. These three definitions are conceptually and mathematically straightforward and we will state them below after introducing some necessary notation and terminology. Some concepts relating to learning theory will be imprecise in what follows to facilitate ease of exposition, but concepts and notation will be made precise where necessary later in this work.

Assume we have variables X and targets Y. Our goal is to learn a function, $f : X \to Y$, that maps inputs to labels. We often denote the predictions from our model as $\hat{y} = f(x)$. We assume that our data is drawn from an underlying distribution (X, Y) and we can interpret our classifier as a random variable at the population level $\hat{Y} = f(X)$. This allows us to reason about the joint distribution $(X, Y, \hat{Y})$.

Classification problems can be binary or multi-class, but we restrict our treatment to the binary case here. In the case of binary classification we can summarize the results of our predictions in a confusion matrix, which contains the information in

Table 2.1 below. For binary classification the target variable, Y, takes on values in $\{0, 1\}$, with $Y = 1$ considered the positive class and $Y = 0$ the negative class.

Predicted	Actual	$P(Predicted \mid Actual)$	Calculation
$\hat{Y} = 1$	$Y = 1$	True positive rate	$\frac{TP}{TP+FN}$
$\hat{Y} = 0$	$Y = 1$	False negative rate	$\frac{FN}{FN+TP}$
$\hat{Y} = 1$	$Y = 0$	False positive rate	$\frac{FP}{FP+TN}$
$\hat{Y} = 0$	$Y = 0$	True negative rate	$\frac{TN}{TN+FP}$

Table 2.1: Information contained in a confusion matrix.

Fairness definitions for classification assume that we have we have "protected attributes" or "sensitive characteristics" which we represent with a discrete random variable A. The features, X, will often implicitly encode information that is highly correlated with these sensitive characteristics, so the naive notion of achieving fairness by merely not passing A to the ML model fails in practice. For instance, web traffic to certain websites is often highly correlated with gender and in many American cities zip codes are highly correlated with race.

Classifiers often work by generating a "score" and then thresholding this score to produce a prediction of the target label. A natural score function is the expected value of Y given X, *i.e.* $R = \mathbb{E}[Y \mid X = x]$. The majority of proposed fairness criteria are properties of the joint distribution of A, Y, and R. The three fairness definitions given above are defined in Table 2.2, with $\perp$ denoting independence of random variables. Separation, or a relaxation of it known as equalized opportunity, is arguably the most commonly used definition of fairness. For binary classification, separation is satisfied if protected groups have the same false positive and false negative rates. That is, for $Y = \{0, 1\}$ and $A = \{a, b\}$ separation is equivalent to

$$P(R = 1 \mid Y = 1, A = a) = P(R = 1 \mid Y = 1, A = b)$$
$$P(R = 1 \mid Y = 0, A = a) = P(R = 1 \mid Y = 0, A = b).$$

Fairness Name	Definition
Independence	$R \perp A$
Separation	$R \perp A \mid Y$
Sufficiency	$Y \perp A \mid R$

Table 2.2: Most common fairness criteria.

Due to the novelty of research into fairness, there are a large number of fairness criteria developed by researchers that are either equivalent, or relaxations of these three criteria. For instance, independence is also known as statistical parity, group fairness, and demographic parity. The mathematical definitions do not necessarily make clear the motivation for the fairness definitions, so we provide the intuitive explanation for what these definitions mean below.

- **Independence**: A classifier is fair if the risk scores generated by the classifier are independent of the sensitive characteristic. That is, the classifier's output must not be correlated with the sensitive characteristic.

- **Separation**: A classifier is fair if the risk scores generated by the classifier are independent of the sensitive characteristic, conditioned on the true labels. That is, the classifier's output can be correlated with the sensitive characteristic only to the extent that is justified by the true labels. If a classifier's output is correlated with the sensitive characteristic after conditioning on the true label, then that classifier is unfair.

- **Sufficiency**: The least intuitive of the three definitions, sufficiency is related to the notion of *calibration*. A classifier is fair if the probability of being assigned to the positive class is the same across groups given a risk score $R = r$.

It is not hard to come up with examples of contexts in which each of these three definitions, or their many related relaxations, do not provide an adequate definition

of fairness. For instance, an example borrowed from Davies *et al.* [5] regarding recidivism and gender reveals a scenario in which satisfying independence actually results in an unfair classifier. Female inmates tend to recidivate at a lower rate than males and an algorithm that decides whether or not to grant parole to inmates that satisfies independence thus discriminates unfairly against female inmates. In fact, any scenario in which there is legitimate correlation between the sensitive characteristic and the target variable is inappropriate for the independence definition of fairness. Conversely, there are many scenarios where data will have illegitimate correlation between sensitive attributes and target variables due to historical injustices. In these scenarios, training a classifier that adheres to the separation definition would likely result in an unfair model.

These definitions are problematic in a variety of ways, but represent an attempt to formally define the notion of fairness in ML. The question of which definition is appropriate depends on the context and a variety of non-mathematical issues. The definitions are also not equivalent. In fact, it is relatively trivial to show that they are incompatible. In almost all realistic scenarios it is impossible to satisfy any two of these fairness criteria simultaneously [2].

2.1.1 Issues With Fairness Definitions

We briefly outline four issues with the kinds of fairness definitions we discussed above. Attempts to address and better understand issues 2, 3, and 4 will be the subject of this thesis.

1. Which definition should we use? As discussed, this depends on context and often corresponds to specific political or world views.

2. The definitions apply to static supervised learning problems. The world, however, is not static. Distributions upon which we train and deploy models shift, often in response to the very model that is deployed. Static supervised learning

with independent and identically distributed data is a dangerous simplification of the world in which we deploy high stakes ML models.

3. We often do not have access to demographic information. All of these definitions require having access to the sensitive characteristic, A. We cannot use these definitions to ensure we have a non-discriminatory model if we do not have access to this sensitive group data.

4. They ignore intersectionality. Preventing discrimination against protected subgroups does not prevent discrimination against vulnerable populations that lie at the intersection of these subgroups, but do not have additional protected status.

If, given a specific context and machine learning problem, we are able to agree on what the appropriate definition of fairness is, we can potentially use our formal non-discrimination criteria, or fairness definitions, to train a model that we consider to be fair. Even when we can do this though, we need to be able to identify exactly which characteristics we consider to be sensitive and those characteristics must be reliably present in the data. If all of these things are true, then there are a variety of methods to ensure we train a fair algorithm including data pre-processing, constrained optimization at training time, or post-processing after an algorithm has been trained.

There are of course many scenarios where those assumptions are not satisfied and we are unable to ensure that our classifier is fair with regards to a formal definition.

All of these approaches rely on having information about the sensitive attribute or characteristic contained in the data, however, and there are many scenarios where we simply do not have sensitive demographic information. There are in fact many scenarios where collection of sensitive demographic information is explicitly forbidden in order to attempt to prevent discrimination based on these characteristics. In these scenarios our formal non-discrimination criteria do not offer us a method to ensure we train fair classifiers and alternative methods for achieving fairness are required.

2.2 Empirical Risk Minimization

In supervised learning problems we make an assumption that our data is generated from a data generating process. This data generating process can be characterized by a probability distribution, allowing us to use the tools of probability theory to reason about the learning problem. Any finite dataset is (generally) assumed to be sampled independently and identically distributed (i.i.d.) from this data distribution and the goal is to learn a model that can approximate this data generating distribution. We only ever have access to finite data (*i.e.* samples from our data generating distribution), however, and we use these samples to learn a model. In order to reason about how a model will work in practice we want a measure of "risk"[1].

In the supervised learning problem we have two spaces of objects X and Y and we want to learn a function $h : X \to Y$ known as a "hypothesis". We assume a joint distribution $P(x, y)$ over X and Y and assume training samples, (x_i, y_i) are drawn i.i.d. from $P(x, y)$. We model y as a random variable with conditional distribution $P(y|x)$ for a fixed x. We are given a non-negative loss function $\ell(\hat{y}, y)$ which gives a measure of difference of prediction $\hat{y}$ from a hypothesis and the true outcome y. The *risk* associated with $h(x)$ is defined as:

$$R(h) = \mathbb{E}[\ell(h(x), y)] = \int \ell(h(x), y) dP(x, y)$$

The goal of the learning algorithm is to find a hypothesis $h^* \in \mathcal{H}$ such that

$$h^* = \arg\min_{h \in \mathcal{H}} R(h).$$

As mentioned above, however, we don't have access to $P(x, y)$ so we can't compute $R(h)$. Instead, we can compute an approximation which we call *empirical risk*

$$R_{emp}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i).$$

[1]Note that our description of risk minimization in what follows is arguably a limited perspective that more recent works such as Bartlett and Anthony [12] expand upon, but we sacrifice some level of generality and rigor here in favour of ease of exposition.

Hence, the learning algorithm should choose a hypothesis $\hat{h}$ such that

$$\hat{h} = \arg\min_{h \in \mathcal{H}} R_{emp}(h).$$

This process is known as *empirical risk minimization* (ERM) and is the de facto training process used in most supervised learning scenarios. ERM is intuitively appealing and has important theoretical guarantees associated with it [13]. It can, however, be problematic when it comes to fairness concerns.

Since we are averaging the loss over our data points, in general, ERM causes an algorithm to focus on majority cases while ignoring minority cases or rare events. For instance, consider a population composed of two sub-populations characterized by differing probability distributions. Assume 80% of the population comes from sub-population 1 and 20% from sub-population 2. This means that for ERM, 80% of the potential loss lies with sub-population 1, and if a learning algorithm is forced to trade-off performance over the two sub-populations, a model trained through ERM will likely learn to fit sub-population 1 better than sub-population 2. A concrete example of this is facial recognition.

For instance, work by Joy Buolamwini and Timnit Gebru [14] showed that commercially available facial recognition software exhibited high performance for white male faces while it performed very poorly for black female faces. In 2020, a Twitter conflict over Duke University's PULSE AI photo recreation tool that turned non-white faces into generic white faces resulted in Yann LeCun temporarily quitting the platform. Both of these notable controversies of bias related to skin colour were the result of not taking necessary measures to prevent bias when training a model from a biased dataset that contained far more white than non-white faces.

While ERM has desirable mathematical properties and is intuitively appealing, it is discordant with notions of equality and fairness in democratic societies. A general principle of democratic societies is that individuals must receive non-disparate treatment, that is, everyone should be treated equally. For instance, individuals have

equal rights under the law, we invest large sums of money to ensure equal access to spaces for people of different abilities, and it is generally illegal to discriminate against individuals on the basis of particular characteristics.

Relying on ERM as an objective to train ML models does not adhere to this notion of non-disparate treatment. When we use ERM to train algorithms that will make predictions about human beings, we are implicitly deciding to favour whichever group composes the majority sub-probability distribution within the training data. In some contexts, the sub-probability distributions will not be correlated to characteristics or demographics in ways that we feel are undesirable, but in many cases sub-groups within populations are likely to have differing probability distributions on the basis of race, gender, class, sexual orientation, religious beliefs, or many other criteria that are unacceptable to discriminate against in a pluralistic, democratic society. When we see this kind of correlation between probability distributions and sensitive characteristics, we risk training algorithms that perform poorly on minority groups, potentially further entrenching systems of inequality.

The fairness definitions outlined above were developed to combat this, but, as we explained, these definitions are limited in their utility. Due to their incredible efficiency and effectiveness, AI algorithms are likely to continue to be deployed and integrated into our lives and economies at an ever greater pace. In order to ensure that these algorithms make decisions in ways that align with our notions of fairness and non-discrimination we will require objectives that can provide some assurance that we will train non-discriminatory models.

2.3 Distributionally Robust Optimization

The problems with ERM and the fairness definitions outlined above are concerning and there is not an obvious solution to the problems. An alternative approach to ERM, known as distributionally robust optimization (DRO), offers an elegant option that addresses many of these concerns. There are a wide variety of formulations for

DRO, but we follow the approach outlined in Duchi and Namkoong [15].

As opposed to minimizing the average loss, distributionally robust optimization considers the *distributionally robust* problem:

$$\underset{\theta\in\Theta}{\text{minimize}} \left\{ R_f(\theta, P_0) := \sup_{Q \ll P_0} \{\mathbb{E}_Q[\ell(\theta; X)] : D_f(Q||P_0) \leq \rho\} \right\},$$

where $\Theta \subset \mathbb{R}^d$ is the parameter (model) space, P_0 is the data generating distribution on the measure space $(\mathcal{X}, \mathcal{A})$, X is a random element of $\mathcal{X}$ and $\ell : \Theta \times \mathcal{X} \to \mathbb{R}$ is a loss function.

In this formulation of DRO, $D_f(Q||P_0)$ is an f-divergence between Q and P_0 and $\{\mathbb{E}_Q[\ell(\theta; X)] : D_f(Q||P_0) \leq \rho\}$ is the set of all expected losses over the f-divergence ball of radius ρ. Alternative DRO formulations utilizing different measures of distance between probability distributions such as Wasserstein balls have also been explored [16–19]. The notation $Q \ll P_0$ means that Q is absolutely continuous with respect to P_0[2]. An f-divergence, $D_f(Q||P)$, is a function that measures the difference between two probability distributions Q and P, although it is not a metric. There are a variety of different f-divergences, but their general definition is as follows [20–22].

$D_f(Q||P) := \int f(\frac{dQ}{dP})dP$ where f is a convex function such that $f(1) = 0$. If Q and P are absolutely continuous with respect to a reference distribution μ, then their probability densities q and p satisfy $dP = p$ and $dQ = q$ and the f-divergence can be written as

$$D_f(Q||P) = \int f(\frac{q(x)}{p(x)})d\mu(x).$$

Some examples of f-divergences are KL-divergence ($f(t) = t\log t$), Pearson χ^2-divergence ($f(t) = (t-1)^2, t^2-1, t^2-t$), and total variation distance ($f(t) = \frac{1}{2}|t-1|$). An f-divergence ball centred at P of radius ρ is a set that contains all probability distributions whose f-divergence is within ρ of P. For example, $\{Q : D_f(Q||P_0) \leq \rho\}$ is an f-divergence ball that contains all probability distributions Q that are within ρ

[2]If two measures, μ and ν, are on the same measure space $(\mathcal{X}, \mathcal{A})$, μ is said to be absolutely continuous with respect to ν if $\mu(A) = 0$ for every set A for which $\nu(A) = 0$.

f-divergence distance from P_0. This is analogous to the notion of an ϵ-ball surrounding a vector in a vector space.

The distributionally robust problem therefore, is to find a set of parameters, $\theta \in \Theta$, that minimize the worst case expected loss of all probability distributions Q that are within the f-divergence ball of radius ρ of our data generating distribution. This differs from ERM in that we do not seek to minimize the expected loss over our data generating distribution, but instead seek to minimize the worst case expected loss over a set of probability distributions nearby our data generating distribution.

As the name suggests, learning parameters that minimize the distributionally robust risk gives us a model that is robust to changes in the data generating distribution. Traditional risk minimization will return a model that minimizes risk for the data generating distribution, but offers no performance guarantees when that distribution changes, and can result in models that are brittle and do not perform well on out of distribution (OOD) examples. DRO, on the other hand, is a more conservative procedure that minimizes worst-case risk and should thus be robust to changes to the probability generating distribution that lie within the f-divergence ball of radius ρ.

As with traditional risk minimization, we do not have access to the theoretical data generating distribution. Instead, we solve the distributionally robust problem via the plug-in estimator

$$\hat{\theta}_n \in \arg\min_{\theta \in \Theta} \left\{ R_f(\theta, \hat{P}_n) := \sup_{Q \ll \hat{P}_n} \{\mathbb{E}_Q[\ell(\theta; X)] : D_f(Q||\hat{P}_n) \leq \rho\} \right\}$$

where $\hat{P}_n$ is the empirical measure on $X_i \sim^{iid} P_0$. Duchi and Namkoong prove convergence guarantees and rates for the plug-in estimator along with some asymptotic results such as showing consistency of the estimator [15].

It is not necessarily clear from the discussion so far what DRO has to do with fairness concerns in machine learning, but DRO in fact has characteristics that are very desirable for addressing issues of fairness and discrimination in learning algorithms. We will explain this shortly, but first we introduce dual reformulations of

the distributionally robust problem as it provides some intuition as to the connection with fairness concerns.

The following theorem comes from Shapiro [23]. Let $f^*(s) := \sup_t\{st - f(t)\}$ be the Fenchel conjugate.

Theorem 1 *(Shapiro [[23], Section 3.2]). Let P be a probability measure on $(\mathcal{X}, \mathcal{A})$ and $\rho > 0$. Then*

$$R_f(\theta; P) = \inf_{\lambda \geq 0, \eta \in \mathbb{R}} \left\{ \mathbb{E}_P \left[\lambda f^* \left(\frac{\ell(\theta; X) - \eta}{\lambda} \right) \right] + \lambda\rho + \eta \right\}$$

for all θ. Moreover, if the supremum on the left hand side is finite, there are finite $\lambda(\theta) \geq 0$ and $\eta(\theta) \in \mathbb{R}$ attaining the infimum on the right hand side.

Additionally, Duchi and Namkoong [15] provide a simplified version of this dual formulation for the Cressie-Read family of f-divergences, obtained by minimizing out $\lambda > 0$ from Theorem 1.

Theorem 2 *(Duchi and Namkoong [[15], Section 2]). For any probability P on $(\mathcal{X}, \mathcal{A})$, $k \in (1, \infty)$, $k_* = k/(k-1)$, any $\rho > 0$, and $c_k(\rho) = (1 + k(k-1)\rho)^{1/k}$, we have for all $\theta \in \Theta$*

$$R_f(\theta; P) = \inf_{\eta \in \mathbb{R}} \left\{ c_k(\rho) \mathbb{E}_P \left[(\ell(\theta; X) - \eta)_+^{k_*} \right]^{1/k_*} + \eta \right\}$$

The above formulations are jointly convex in (θ, λ, η) and (θ, η), respectively, for convex losses, $\ell(\theta; X)$, making them amenable to techniques from convex optimization such as interior point methods [24].

The Cressie-Read family of f-divergences are parameterized by $k \in (-\infty, \infty) \setminus \{0, 1\}$, $k_* = \frac{k}{k-1}$, with

$$f_k(t) := \frac{t^k - kt + k - 1}{k(k-1)} \quad \text{and} \quad f_k^*(s) := \frac{1}{k} \left[((k-1)s + 1)_+^{k_*} - 1 \right].$$

Issues of fairness arise in machine learning when an algorithm treats different demographic groups in a disparate fashion. As outlined above, this often means that

algorithmic performance varies across different demographic groups. If this occurs, it is almost certainly the case that there exist distinct probability distributions over the demographic groups because if all demographic groups were characterized the the same probability distribution, classification algorithms should exhibit uniform performance across demographic groups.

This realization makes it easy to see why ERM is likely to discriminate, particularly against minority groups. ERM treats the loss on each data point equally, thus, if there is a majority and minority probability distribution and a learning algorithm must balance performance on these groups, an ERM objective is likely to result in a model that performs better on the majority group and worse on the minority group. It is exactly this type of problem that DRO is designed to resolve.

Unlike ERM, DRO does not equally weight each data point, but instead up-weights data points on which the model is achieving high loss. This means that the model should achieve somewhat uniform performance on individuals across demographic groups. This can be seen by looking at the dual formulation in Theorem 2. The DRO objective only considers losses above the optimal dual variable $\eta^*(\theta)$ and these losses are up-weighted by the $L^{k_*}(P)$-norm. Losses that are less than the optimal dual variable are set to zero in the objective. Another way of saying this is that the DRO objective is equivalent to optimizing the tail-performance of a model [15].

If DRO performs poorly on a subset of the data that is correlated with a distinct demographic group, these losses will be up-weighted by the dual formulation of the distributionally robust objective, pushing the model to improve performance on this subset. DRO does not offer any guarantees of uniform performance across demographic groups, but as we increase the value of ρ we will increase the loss incurred on "hard" regions of the data where the model performs poorly. This often comes with a trade-off with performance on the dataset as a whole and we discuss this further in Chapter 4.

DRO has the potential to solve some of the issues with formal fairness criteria

that we outlined earlier. First, and most importantly, DRO does not require any demographic information in order to protect against poor performance on subsets of the data. The distributionally robust objective optimizes tail-performance of a model and does not require any information about sensitive characteristics to do this. Second, and for the same reason, DRO naturally takes into account intersectionality. It minimizes loss on the worst case distribution within the f-divergence ball, hence it naturally achieves somewhat uniform performance across most subsets of the data. Third, as DRO optimizes for adversarial distributions, it should generally be more robust to distribution drift than ERM. Also, as DRO is merely an optimization technique rather than a formal definition of fairness, it is more amenable to use in scenarios that capture more complexity of real world fairness scenarios.

All of these points come with caveats, however. DRO is not explicitly attempting to learn a model that is fair according to the protected attributes for which we believe the model should not discriminate. The worst case distribution for which DRO optimizes may not be correlated with protected attributes at all, and instead may be some other distribution that exists within our data for which we do not believe fairness concerns apply. An example of when this might occur is in the presence of a large number of outliers. DRO may train a model that attempts to minimize loss on these outliers while not achieving uniform performance across demographic groups. We will discuss more of the practical considerations and issues with training a model using DRO in Chapter 4.

A model trained using DRO is also not guaranteed to adhere to any of the definitions of fairness we discussed in Section 2.1. DRO attempts to minimize the worst-case loss over all probability distributions contained in the f-divergence ball surrounding the data generating distribution. This minimax style optimization procedure is related to the philosopher John Rawls' notion of distributive justice [25], but is not likely to adhere to any of the formal fairness criteria discussed above. We avoid entering into the debate of correct notions or definitions of fairness, largely because we

feel that fairness is inherently contextual and as such, no single definition can suffice. We will instead focus on the more general and intuitive fairness properties that DRO exhibits as compared to ERM.

DRO also does not address the issue of static supervised learning likely being an inadequate conceptual framework to capture the complexity of fairness concerns in machine learning. DRO offers an alternative objective for which to optimize, but if we wish to capture the important dynamics of fairness over time, an alternative formalism will be required. A recently developed concept known as performative prediction, as we explain in the next section, attempts to do exactly this.

2.4 Performative Prediction

In supervised learning we assume that our data is sampled i.i.d. from an unknown data generating distribution and that our model is then deployed to make predictions on data that follows this same distribution. In many scenarios, however, the very act of making predictions influences the data on which we wish to make these predictions. That is to say, our models are *performative* and instead of passively describing the world and making predictions about it, they actually induce change in the world.

A simple example of this is predictive policing. In predictive policing we train a model to predict where crimes are likely to occur based on historical data and then deploy more resources to areas where the model predicts crimes are more likely to occur. The increased police patrols and surveillance results in more crimes being detected which might further increase the perceived crime rate in those areas. If this data is then used for future predictions it will result in a shifted distribution of the data as a result of the predictions of the previous model. Another example of performativity is an algorithm that weights different elements of a student's CV such as SAT scores and GPA in order to make college admissions decisions. If the algorithm heavily weights SAT scores, over time it will become apparent to applicants that SAT scores are very important and they will dedicate more resources to improving SAT

scores, thus changing the distribution of the data on which the algorithm is making predictions as a result of those very predictions.

Performative prediction is closely related to many other fields in machine learning including bandits, reinforcement learning, strategic classification, causal inference, convex optimization, and game theory, but the precise notion and formalism for performative prediction was only developed very recently in Perdomo *et al.* [3]. We formally specify the performative prediction problem now and contrast it with the supervised learning problem.

Assume we have a measure space $(\mathcal{Z}, \mathcal{A})$ with Z a random element of $\mathcal{Z}$ and $\mathcal{D}$ the data generating distribution on this space. Let $\Theta \subset \mathbb{R}^d$ be the parameter (model) space and $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$ be a loss function. The supervised learning problem is to minimize the the objective over this distribution. In the case of risk minimization the objective is the expected loss, $\ell(Z; \theta)$ with respect to $\mathcal{D}$.

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

In contrast to this, performative prediction involves making predictions on a distribution that has been shifted as a result of deploying the model, $\mathcal{D}(\theta)$. We refer to $\mathcal{D}(\theta)$ as the *distribution map*. The concept that captures this notion of risk is known as *performative risk* and is formalized as follows:

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

The difference between this and the supervised learning problem is that expected loss is now taken with respect to the induced distribution rather than the data generating distribution.

The notion of what constitutes a good model is different in supervised learning and performative prediction. In supervised learning the task is simpler - minimize the risk on the data generating distribution. In performative prediction however, we now need to consider how to minimize risk on a distribution that is different from that

which generated our training data, and is in fact a function of whatever model we deploy. To capture these notions, Perdomo *et al.* [3] define *performative optimality* and *performative stability.*

Definition 3 *(performative optimality) A model $f_{\theta_{PO}}$ is performatively optimal if the following relationship holds:*

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

Equivalently, $\theta_{PO} = \arg\min_{\theta \in \Theta} PR(\theta)$ where $PR(\theta)$ is the performative risk as defined above.

A performatively optimal point is a minimizer of the performative risk. An alternative solution concept is referred to as *performative stability.*

Definition 4 *(performative stability) A model $f_{\theta_{PS}}$ is performatively stable if the following relationship holds:*

$$\theta_{PS} = \arg\min_{\theta \in \Theta} \mathbb{E}_{Z \sim \mathcal{D}(\theta_{PS})}[\ell(Z; \theta)].$$

Define $DPR(\theta, \theta') := \mathbb{E}_{Z \sim \mathcal{D}(\theta)}[\ell(Z; \theta')]$ as the decoupled performative risk; then $\theta_{PS} = \arg\min_{\theta \in \Theta} DPR(\theta_{PS}, \theta)$.

A performatively stable model is not necessarily a minimizer of the performative risk, but it is optimal on the distribution it induces. Hence, if you have performative stability there is no need to retrain a model to cope with the induced distribution drift. Performative optimality and performative stability are distinct concepts and a performatively optimal point is not necessarily performatively stable, and vice versa. As explained in Perdomo *et al.* [3], performatively stable models are fixed points of risk minimization.

In game theoretic terms, we can consider performative prediction as a game in which one player deploys a model, θ, and the environment responds with some distribution map, $\mathcal{D}(\theta)$. If $\mathcal{D}(\theta)$ is a best response, then a performatively optimal point

corresponds to a Stackelberg equilibrium, whereas a performatively stable point corresponds to a Nash equilibrium. From game theory we know that except in special cases (*e.g.* finite zero-sum games), Nash equilibria and Stackelberg equilibria do not necessarily coincide [26].

Performative prediction presents a special case of learning under distribution drift, where the distribution drift is a function of the model deployed. A common approach in supervised learning under distribution drift is to retrain a model on newly collected data. While this does not directly minimize performative risk, it is a potentially reasonable solution in a variety of scenarios. Perdomo *et al.* [3] prove theorems under which repeated risk minimization and repeated gradient descent converge to performatively stable models. We present these findings in detail here as they are relevant for our work.

Definition 5 *(ϵ-sensitivity) We say that a distribution map $\mathcal{D}(\cdot)$ is ϵ-sensitive if for all $\theta, \theta' \in \Theta$:*

$$W_1\left(\mathcal{D}(\theta), \mathcal{D}(\theta')\right) \leq \epsilon ||\theta - \theta'||_2,$$

where W_1 denotes the Wasserstein-1 distance.

We also make assumptions on the loss function $\ell(z;\theta)$. Let $\mathcal{Z} := \cup_{\theta\in\Theta} supp(\mathcal{D}(\theta))$.

Definition 6 *(joint smoothness) We say that a loss function $\ell(z;\theta)$ is β-jointly smooth if the gradient ∇_θ is β-Lipschitz in θ and z, that is*

$$||\nabla_\theta \ell(z;\theta) - \nabla_\theta \ell(z;\theta')||_2 \leq \beta ||\theta - \theta'||_2, \quad ||\nabla_\theta \ell(z;\theta) - \nabla_\theta \ell(z';\theta)||_2 \leq \beta ||z - z'||_2,$$

for all $\theta, \theta' \in \Theta$ and $z, z' \in \mathcal{Z}$

Definition 7 *(strong convexity) We say that a loss function $\ell(z;\theta)$ is γ-strongly convex if*

$$\ell(z;\theta) \geq \ell(z;\theta') + \nabla_\theta \ell(z;\theta')^T(\theta - \theta') + \frac{\gamma}{2}||\theta - \theta'||_2^2,$$

for all $\theta, \theta' \in \Theta$ and $z \in \mathcal{Z}$. If $\gamma = 0$ this condition is equivalent to convexity.

Definition 8 *(repeated risk minimization) Repeated risk minimization (RRM) refers to the procedure where, starting from an initial model f_{θ_0}, we perform the following sequence of updates for every $t \geq 0$:*

$$\theta_{t+1} = G(\theta_t) := \arg\min_{\theta \in \Theta} \mathbb{E}_{Z \sim \mathcal{D}(\theta_t)}[\ell(Z; \theta)].$$

Definitions 5, 6, and 7 are assumptions on the loss and distribution map that are required for Theorem 3.5 in Perdomo *et al.* [3]. We state that theorem now.

Theorem 9 *(Perdomo et al. [[3], Theorem 3.5]) Suppose that the loss $\ell(z; \theta)$ is β-jointly smooth and γ-strongly convex. If the distribution map $\mathcal{D}(\cdot)$ is ϵ-sensitive, then the following statements are true:*

1. *$||G(\theta) - G(\theta')||_2 \leq \epsilon\frac{\beta}{\gamma}||\theta - \theta'||_2$, for all $\theta, \theta' \in \Theta$*
2. *If $\epsilon < \frac{\gamma}{\beta}$, the iterates θ_t of RRM converge to a uniquely performatively stable point θ_{PS} at a linear rate: $||\theta_t - \theta_{PS}||_2 \leq \delta$ for $t \geq \left(1 - \epsilon\frac{\beta}{\gamma}\right)^{-1} \log\left(\frac{||\theta_0 - \theta_{PS}||_2}{\delta}\right)$*

This theorem says that with the appropriate smoothness and convexity conditions satisfied on the loss and distribution map, repeatedly retraining a model by minimizing risk will converge to a unique performatively stable model at a linear rate. The proof proceeds by demonstrating that under the assumptions stated in Theorem 9, the RRM operator is a contraction mapping. Perdomo *et al.* [3] also show that without any of ϵ-sensitivty, β-joint smoothness, or γ-strong convexity one can produce examples that do not converge to a fixed point. Hence these assumptions are necessary conditions to guarantee convergence of RRM to a performatively stable model in the general case.

In addition to Theorem 3.5, Perdomo *et al.* show that repeated gradient descent (RGD) also converges to a performatively stable model. We define RGD and state that theorem here.

Definition 10 *(repeated gradient descent) Repeated gradient descent (RGD) is the procedure where, starting from an initial model f_{θ_0}, we perform the following sequence of updates for every $t \geq 0$:*

$$\theta_{t+1} = G_{gd}(\theta_t) := \Pi_\Theta \left(\theta_t - \eta \mathbb{E}_{Z \sim \mathcal{D}(\theta_t)} [\nabla_\theta \ell(Z; \theta_t)] \right),$$

where $\eta > 0$ is a fixed step size and Π_Θ denotes the Euclidean projection operator onto Θ.

Theorem 11 *(Perdomo et al. [10, Theorem 3.8]) Suppose that the loss $\ell(z; \theta)$ is β-jointly smooth and γ-strongly convex. If the distribution map $\mathcal{D}(\cdot)$ is ϵ-sensitive with $\epsilon < \frac{\gamma}{(\beta+\gamma)(1+1.5\eta\beta)}$, then RGD with step size $\eta \leq \frac{2}{\beta+\gamma}$ satisfies the following:*

1. $||G_{gd}(\theta) - G_{gd}(\theta')||_2 \leq \left(1 - \eta \left(\frac{\beta\gamma}{\beta+\gamma} - \epsilon(1.5\eta\beta^2 + \beta)\right)\right) ||\theta - \theta'||_2 < ||\theta - \theta'||$
2. *The iterates θ_t of RGD converge to a uniquely performatively stable point θ_{PS} at a linear rate: $||\theta_t - \theta_{PS}||_2 \leq \delta$ for $t \geq \frac{1}{\eta} \left(\frac{\beta\gamma}{\beta+\gamma} - \epsilon(1.5\eta\beta^2 + \beta)\right)^{-1} \log\left(\frac{||\theta_0 - \theta_{PS}||_2}{\delta}\right)$*

Unlike for Theorem 9, Perdomo *et al.* [3] do not show that assumptions 5, 6, and 7 are necessary conditions for Theorem 11 to hold.

Interestingly, while strong convexity often results in faster convergence rates for supervised learning problems, it is not usually a necessary property for convergence. In the performative prediction framework, however, strong convexity, rather than merely convexity, is required to guarantee convergence. A series of papers following Perdomo *et al.* [3] have further expanded upon the theory of performative prediction [27–30].

2.5 Related Work

Despite the relative recency of the development of the formal fairness criteria described in Section 2.1, there have been attempts to address some of the shortcomings of these static supervised learning fairness definitions in the past couple of years.

Many of these works provided the inspiration for this thesis, so we briefly highlight some of this related work.

The two most influential areas of work are performative prediction and distributionally robust optimization, and the papers most important for this work are Perdomo *et al.* [3] and Duchi and Namkoong [15]. Perdomo *et al.* develop the performative prediction framework, while Duchi and Namkoong give a thorough summary of distributionally robust optimization and explain its potential application to fairness concerns in machine learning.

Following the publication of *Performative Prediction* [3], there have been a number of papers which have further explored performative prediction and extended some of the results from the original paper. For instance, Mendler-Dunner *et al.* [27] prove results for stochastic optimization in performative prediction, Miller *et al.* [28] prove new results relating performatively stable points to performatively optimal points, Brown *et al.* [29] attempt to move toward sequential games for performative prediction by adding a notion of state to the performative prediction problem, and Dong and Ratliff [30] approach performative prediction from a dynamical systems perspective allowing a move away from strict contraction mappings when examining convergence to performatively stable models.

Distributionally robust optimization is an older area of research as compared to performative prediction, but it has only recently received more attention within the machine learning community, largely due to work by Hongseook Namkoong and John Duchi [31–33]. Areas such as finance, where robust optimization is important as rare events have the potential to be catastrophic for a portfolio, have seen research on robust optimization for decades. See Ben-Tal *et al.* [34] for a survey. Ben-Tal *et al.* [35] and Shapiro [23] provide rigorous mathematical treatments of DRO with f-divergence balls.

A number of recent papers have also attempted to better understand fairness in machine learning over time, albeit not within the performative prediction framework.

Closely related to performative prediction, strategic classification [36] models the prediction problem as an adversarial game where individuals manipulate their features in response to deployed models. Recent work [37, 38] has examined how strategic classification interacts with fairness concerns. Strategic classification is an interesting problem setting, but is not as general as performative prediction and only captures scenarios that can rightly be modelled as adversarial games. D'Amour *et al.* [39] on the other hand, approach long-term fairness questions from a purely empirical perspective and use simulation studies to understand the long-term dynamics of algorithmic choices on populations in a variety of scenarios designed to reflect real-world applications. This approach is interesting, but the lack of theoretical framework limits their ability to generalize beyond specific simulations.

Finally, the work most closely related to this thesis, Hashimoto *et al.* [40], examines the impact of repeatedly minimizing loss and deploying a model using empirical risk minimization and distributionally robust optimization on a population that changes as a function of the loss incurred. This work, however, does not fit into the performative prediction framework and applies only to a specific scenario in which individuals arrive according to a Poisson process and depart as a function of the loss. This thesis aims to extend this work by combining DRO and performative prediction in order to leverage the known results in both areas to provide a more general understanding of how DRO can impact long-term fairness.

Chapter 3

Performative Prediction and Distributionally Robust Optimization: Performative DRO

Up to this point it is not necessarily immediately clear how fairness criteria, empirical risk minimization, distributionally robust optimization, and performative prediction are related, but we will explain the connection now.

As discussed earlier, the formal non-discrimination criteria that have been the focus of much research of fairness in ML are really only well-defined for a static supervised learning scenario. Many, if not most, real world scenarios involving fairness concerns cannot be characterized as static supervised learning problems. The world is in a constant state of flux, and many distribution changes are in fact the result of deploying models to make predictions about the world. Performative prediction therefore presents a framework in which we can reason about the fairness properties of algorithms that is closer to our real world concerns than the static supervised learning framework.

The two examples given above, predictive policing and college admissions, clearly have serious fairness concerns associated with them. Some other common examples of fairness scenarios which involve performativity are parole decisions, credit extension decisions, and hiring decisions. Given the importance of performativity for fairness, our fairness definitions for static distributions seem even more unsatisfying and prob-

lematic.

Research on performative prediction, however, has thus far focused on risk minimization (*i.e.* expected losses). As we noted in Section 2.3, training models with a risk minimization objective in the presence heterogeneous populations characterized by distinct probability distributions has the potential to result in algorithms that discriminate against minority groups. The result is that performatively stable and performatively optimal models are likely to be unfair models.

Understanding the dynamics of performative prediction under alternative objectives that prioritize fairness is an important, and as yet unexplored, area of research for fairness in ML. If we believe that many real-world scenarios with fairness considerations involve performative prediction, it is essential that we have a better understanding what fairness looks like in the presence of performative prediction.

In Section 2.2 we introduced risk minimization and empirical risk minimization and in Section 2.4 we discussed performative prediction. In this chapter we now consider replacing the risk minimization objective in performative prediction with a distributionally robust objective. We will proceed by adapting definitions, assumptions, and theorems from Perdomo *et al.* [3] to a distributionally robust objective.

3.1 Definitions for Performative DRO

We begin by redefining performative risk and performatively optimal and stable models in terms of a robust objective. In the following $\mathcal{D}(\theta)$ is a distribution map and $D_f(Q||\mathcal{D}(\theta))$ is an f-divergence ball of radius ρ.

Definition 12 *(robust performative risk) The robust performative risk of a model, θ, is*

$$RPR(\theta) = \sup_{Q \ll \mathcal{D}(\theta)} \{\mathbb{E}_{Z\sim Q}[\ell(Z;\theta)] : D_f(Q||\mathcal{D}(\theta)) \leq \rho\}.$$

Robust performative risk differs from performative risk in that the induced distribution, $\mathcal{D}(\theta)$, is now the centre of an f-divergence ball over which a supremum of the

expected loss is taken.

Definition 13 *(robust performative optimality) A model $f_{\theta_{PO}}$ is robustly performatively optimal if the following relationship holds:*

$$\theta_{PO} = \arg\min_{\theta\in\Theta} \sup_{Q\ll\mathcal{D}(\theta)} \{\mathbb{E}_{Z\sim Q}[\ell(Z;\theta)] : D_f(Q||\mathcal{D}(\theta)) \leq \rho\}.$$

Equivalently, $\theta_{PO} = \arg\min_{\theta\in\Theta} RPR(\theta)$ where $RPR(\theta)$ is the robust performative risk as defined above.

Definition 14 *(robust performative stability) A model $f_{\theta_{PS}}$ is robustly performatively stable if the following relationship holds:*

$$\theta_{PS} = \arg\min_{\theta\in\Theta} \sup_{Q\ll\mathcal{D}(\theta_{PS})} \{\mathbb{E}_{Z\sim Q}[\ell(Z;\theta)] : D_f(Q||\mathcal{D}(\theta_{PS})) \leq \rho\}.$$

Define $RDPR(\theta,\theta') := \sup_{Q\ll\mathcal{D}(\theta)}\{\mathbb{E}_{Z\sim Q}[\ell(Z;\theta')] : D_f(Q||\mathcal{D}(\theta)) \leq \rho\}$ as the robust decoupled performative risk; then $\theta_{PS} = \arg\min_{\theta\in\Theta} RDPR(\theta_{PS},\theta)$.

We discussed earlier that previous work has shown that repeated risk minimization converges to a performatively stable model under certain assumptions on the loss and distribution map [3]. We can analogously define repeated distributionally robust optimization as follows.

Definition 15 *(repeated distributionally robust optimization) Repeated distributionally robust optimization (RDRO) refers to the procedure where, starting from an initial model f_{θ_0}, we perform the following sequence of updates for every $t \geq 0$:*

$$\theta_{t+1} = G(\theta_t) := \arg\min_{\theta\in\Theta} \sup_{Q\ll\mathcal{D}(\theta_t)} \{\mathbb{E}_{Z\sim Q}[\ell(Z;\theta)] : D_f(Q||\mathcal{D}(\theta_t)) \leq \rho\}.$$

As with RRM, this is an iterative procedure where we optimize the distributionally robust objective at each time step t. While it is only a small departure from RRM conceptually, the DRO objective has fundamentally different mathematical properties than risk minimization making it unclear whether RDRO and RRM should exhibit similar behaviour.

We show that by modifying the assumptions required for convergence of RRM, we can use the proof technique from Perdomo *et al.* [3] to show that RDRO converges to a robustly performatively stable model. We now state and discuss these modified assumptions.

Definition 16 *(robust β-joint smoothness) We say the the distributionally robust objective is robustly β-jointly smooth if for all $\theta, \theta' \in \Theta$ and $z, z' \in \mathcal{Z}$ the gradient,*

$$\nabla_\theta \sup_{Q \ll \mathcal{D}(\theta)} \{\mathbb{E}_{Z\sim Q}[\ell(Z;\theta)] : D_f(Q||\mathcal{D}(\theta)) \leq \rho\},$$

exists and is β-Lipschitz in z and θ.

Definition 17 *(robust γ-strong convexity) We say the the distributionally robust objective is robustly γ-strongly convex if for all $\theta, \theta' \in \Theta$ and $z \in \mathcal{Z}$*

$$\sup_{Q \ll \mathcal{D}(\theta)} \{\mathbb{E}_{Z\sim Q}[\ell(Z;\theta)] : D_f(Q||\mathcal{D}(\theta)) \leq \rho\}$$

is γ-strongly convex.

Definition 18 *(robust ϵ-sensitivity)*
Let $\mathcal{D}^*(\theta) = \arg\max_{Q: D_f(Q||\mathcal{D}(\theta)) \leq \rho} \mathbb{E}_{Z\sim Q}[\ell(Z;\theta)]$. Assume that a distribution map, $D(\cdot)$, is ϵ-sensitive. We say that this distribution map is robustly ϵ-sensitive if there exists $\omega > 0$ such that for any $\theta, \theta' \in \Theta$

$$W_1\left(\mathcal{D}^*(\theta), \mathcal{D}^*(\theta')\right) \leq \omega W_1\left(\mathcal{D}(\theta), \mathcal{D}(\theta')\right) \leq \omega\epsilon||\theta - \theta'||_2,$$

These definitions are straightforward extensions of β-joint smoothness, γ-strong convexity, and ϵ-sensitivity to the distributionally robust objective. With risk minimization it is enough to make the assumptions of β-joint smoothness and γ-strong convexity on the loss function, since taking an expectation over the loss preserves smoothness and convexity. This is not the case with DRO, however. When we take the supremum of the expected loss over the f-divergence ball, there is no guarantee that β-joint smoothness will be preserved. In fact, we are not even guaranteed to get

a function that is differentiable everywhere on its domain. A simple example that illustrates this is taking the supremum over two quadratic functions that intersect at a point. The resulting function will not be differentiable at the point of intersection.

Similarly, a supremum over a set of strongly convex functions does not necessarily preserve strong convexity. It is true that the supremum over a set of convex functions preserves convexity, but this does not necessarily extend to strong convexity. Finally, a distribution map that is ϵ-sensitive is not necessarily ϵ-sensitive for worst case distributions, or what we call *robustly ϵ-sensitive*. These complexities make the proof of convergence of repeated DRO significantly more challenging than repeated risk minimization. We do not address these complexities in our proof, but rather we show that if the distributionally robust objective satisfies these properties we get convergence of RDRO analogous to the convergence of RRM. We provide further discussion of these issues after the statement and proof of our theorem.

3.2 A Convergence Theorem for Performative DRO

We will now state and prove a theorem which adapts Theorem 9 from Perdomo *et al.* [3] to RDRO in place of RRM. The proof of this theorem is a straightforward adaptation of the proof in Perdomo *et al.* We first introduce two lemmas used in the proof of Theorem 9 in Perdomo *et al.* [3] which we will make use of in our proof.

Lemma 19 *(First-order optimality condition) Let f be convex and let Ω be a closed convex set on which f is differentiable, then*

$$x_* \in \arg\min_{x \in \Omega} f(x)$$

if and only if

$$\nabla f(x_*)^T (y - x_*) \geq 0, \forall y \in \Omega.$$

Lemma 20 *(Kantorovich-Rubinstein) A distribution map $\mathcal{D}(\cdot)$ is robustly ϵ-sensitive*

if and only if for all $\theta, \theta' \in \Theta$*:*

$$\sup\left\{\left|\mathbb{E}_{Z\sim\mathcal{D}^*(\theta)}[g(Z)] - \mathbb{E}_{Z\sim\mathcal{D}^*(\theta')}[g(Z)]\right| : g : \mathbb{R}^P \to \mathbb{R}, g \text{ 1-Lipschitz}\right\} \leq \omega\epsilon||\theta - \theta'||_2$$

where $\mathcal{D}^*(\theta) = \underset{Q:D_f(Q||\mathcal{D}(\theta))\leq\rho}{\arg\max} \mathbb{E}_{Z\sim Q}[\ell(Z;\theta)]$.

We now state our theorem.

Theorem 21 *Suppose that the distributionally robust objective satisfies definitions 16 and 17, and that the distribution map satisfies definition 18. Then the following statements are true:*

1. $||G(\theta) - G(\theta')||_2 \leq \omega\epsilon\frac{\beta}{\gamma}||\theta - \theta'||_2$*, for all* $\theta, \theta' \in \Theta$
2. *If* $\omega\epsilon < \frac{\gamma}{\beta}$*, the iterates* θ_t *of RDRO converge to a unique robustly performatively stable point* θ_{PS} *at a linear rate:* $||\theta_t - \theta_{PS}||_2 \leq \delta$ *for* $t \geq \left(1 - \omega\epsilon\frac{\beta}{\gamma}\right)^{-1}\log\left(\frac{||\theta_0 - \theta_{PS}||_2}{\delta}\right)$.

Note that $G(\theta) := \arg\min_{\theta\in\Theta}\sup_{Q\ll\mathcal{D}(\theta)}\{\mathbb{E}_{Z\sim Q}[\ell(Z;\theta)] : D_f(Q||\mathcal{D}(\theta)) \leq \rho\}$.

Proof. Fix $\theta, \theta' \in \Theta$. Let

$$\mathcal{D}^*(\theta) = \underset{Q:D_f(Q||\mathcal{D}(\theta))\leq\rho}{\arg\max} \mathbb{E}_{Z\sim Q}[\ell(Z;\theta)].$$

That is, $\mathcal{D}^*(\theta)$ is the distribution within the f-divergence ball centred at $\mathcal{D}(\theta)$ with radius ρ that maximizes the expected loss. Further, let

$$f(\xi) = \mathbb{E}_{Z\sim\mathcal{D}^*(\theta)}[\ell(Z;\xi)] \quad \text{and} \quad f'(\xi) = \mathbb{E}_{Z\sim\mathcal{D}^*(\theta')}[\ell(Z;\xi)].$$

That is, $f(\xi)$ and $f'(\xi)$ are the worst case losses for f-divergence balls centred at θ and θ' respectively with radius ρ.

We now use Definition 17 and Definition 16 to ensure that f is γ-strongly convex and that the gradient of f exists and is β-jointly smooth. With our assumption of γ-strong convexity of f, $G(\theta)$ is the unique minimizer of $f(x)$ and we have the following two inequalities

$$f(G(\theta)) - f(G(\theta')) \geq (G(\theta) - G(\theta'))^T\nabla f(G(\theta')) + \frac{\gamma}{2}||G(\theta) - G(\theta')||_2^2 \tag{3.1}$$

$$f(G(\theta')) - f(G(\theta)) \geq \frac{\gamma}{2}||G(\theta) - G(\theta')||_2^2 \tag{3.2}$$

Inequality (3.1) comes from the definition of γ-strong convexity and inequality (3.2) comes from γ-strong convexity and the first-order optimality condition since Lemma 19 tells us $(G(\theta') - G(\theta))^T \nabla f(G(\theta)) \geq 0$ because $G(\theta)$ is the minimizer of $f(x)$.

Using these two inequalities we can derive the following

$$\begin{aligned} -\gamma||G(\theta) - G(\theta')||_2^2 &\geq f(G(\theta)) - f(G(\theta')) - \frac{\gamma}{2}||G(\theta) - G(\theta')||_2^2 \\ &\geq (G(\theta) - G(\theta'))^T \nabla f(G(\theta')) \end{aligned}$$

where we get the first inequality from (3.2) and the second from (3.1).

Now, using β-joint smoothness in z and Cauchy-Schwarz we get

$$\begin{aligned} ||(G(\theta) - G(\theta'))^T \nabla_\theta \ell(z; G(\theta')) - (G(\theta) - G(\theta'))^T \nabla_\theta \ell(z'; G(\theta'))||_2 \\ \leq ||G(\theta) - G(\theta')||_2 \beta ||z - z'||_2 \end{aligned}$$

That is, $(G(\theta) - G(\theta'))^T \nabla_\theta \ell(z; G(\theta')$ is $||G(\theta) - G(\theta')||_2 \beta$-Lipschitz in z. We will now use this and Kantorovich-Rubinstein (Lemma 20). Let

$$g(z) = \frac{(G(\theta) - G(\theta'))^T \nabla_\theta \ell(z; G(\theta'))}{||G(\theta) - G(\theta')||_2 \beta}$$

The function $g(z)$ is 1-Lipschitz in z because we have just divided $(G(\theta) - G(\theta'))^T \nabla_\theta \ell(z; G(\theta'))$ by its Lipschitz constant. From Lemma 20 and Definition 18 we have the following, with $g(Z)$ as defined above

$$\mathbb{E}_{Z \sim \mathcal{D}^*(\theta)}[g(Z)] - \mathbb{E}_{Z \sim \mathcal{D}^*(\theta')}[g(Z)] \leq \omega\epsilon||\theta - \theta'||_2$$

Using linearity of expectation and multiplying by $||G(\theta) - G(\theta')||_2 \beta$ we get the following

$$\begin{aligned} (G(\theta) - G(\theta'))^T \nabla f(G(\theta')) - (G(\theta) - G(\theta'))^T \nabla f'(G(\theta')) \\ \geq -\omega\epsilon\beta||G(\theta) - G(\theta')||_2 ||\theta - \theta'||_2 \end{aligned}$$

Now again using Lemma 19, we have $(G(\theta) - G(\theta'))^T \nabla f'(G(\theta')) \geq 0$, hence $(G(\theta) - G(\theta'))^T \nabla f(G(\theta')) \geq \omega\epsilon\beta||G(\theta) - G(\theta')||_2 ||\theta - \theta'||_2$. From our work above we showed

that $-\gamma||G(\theta) - G(\theta')||_2^2 \geq (G(\theta) - G(\theta'))^T \nabla f(G(\theta'))$. Putting this all together we get

$$-\gamma||G(\theta) - G(\theta')||_2^2 \geq -\omega\epsilon\beta||G(\theta) - G(\theta')||_2||\theta - \theta'||_2$$

We rearrange the above to get

$$||G(\theta) - G(\theta')||_2 \leq \omega\epsilon\frac{\beta}{\gamma}||\theta - \theta'||_2$$

which proves claim (1) of the theorem.

Claim (2) follows easily. We note that $\theta_t = G(\theta_{t-1})$ from the definition of RDRO and $G(\theta_{PS}) = \theta_{PS}$ by the definition of robust performative stability. Using the result of claim (1) we get

$$||\theta_t - \theta_{PS}||_2 \leq \omega\epsilon\frac{\beta}{\gamma}||\theta_{t-1} - \theta_{PS}||_2 \leq \left(\omega\epsilon\frac{\beta}{\gamma}\right)^t ||\theta_0 - \theta_{PS}||_2$$

Now we set

$$\left(\omega\epsilon\frac{\beta}{\gamma}\right)^t ||\theta_0 - \theta_{PS}||_2 \leq \delta$$

and solve for t.

$$\begin{aligned}
\left(\omega\epsilon\frac{\beta}{\gamma}\right)^t ||\theta_0 - \theta_{PS}||_2 &\leq \delta \\
t\log\left(\omega\epsilon\frac{\beta}{\gamma}\right) + \log(||\theta_0 - \theta_{PS}||_2) &\leq \log(\delta) \\
t\log\left(\omega\epsilon\frac{\beta}{\gamma}\right) &\leq \log(\delta) - \log(||\theta_0 - \theta_{PS}||_2) \\
t\log\left(\omega\epsilon\frac{\beta}{\gamma}\right) \leq t\left(\omega\epsilon\frac{\beta}{\gamma} - 1\right) &\leq \log(\delta) - \log(||\theta_0 - \theta_{PS}||_2) \\
t &\geq (\log(\delta) - \log(||\theta_0 - \theta_{PS}||_2))\left(\omega\epsilon\frac{\beta}{\gamma} - 1\right)^{-1} \\
t &\geq \left(1 - \omega\epsilon\frac{\beta}{\gamma}\right)^{-1} \log\left(\frac{||\theta_0 - \theta_{PS}||_2}{\delta}\right)
\end{aligned}$$

Note that this theorem is essentially just an application of the Banach fixed point theorem as $\omega\epsilon < \frac{\gamma}{\beta} \implies \omega\epsilon\frac{\beta}{\gamma} < 1$. ■

While this proof provides us with conditions under which RDRO converges, it is somewhat unsatisfying as it leaves as an open question what is required for Definitions 16, 17, and 18 to be true. Unlike for performative prediction with risk minimization, making assumptions on our loss function does not guarantee that the distributionally robust objective will share those properties. In order to have smoothness of our objective and also robust ϵ-sensitivity of the distribution map we require some sort of stability or regularity of the worst case distributions as the centre of the f-divergence ball moves as a function of θ.

The robust ϵ-sensitivity condition could possibly be equivalently expressed as a smoothness condition for the f-divergence ball surrounding the data generating distribution. Further exploration of these questions would likely yield further understanding of the distributionally objective, even outside of the context of performative prediction. While intriguing, this work is beyond the scope of this thesis.

It could also be possible that robust β-joint smoothness is not a necessary condition for convergence of RDRO and that a convergence proof may be possible working with subgradients or directional derivatives, but if this is the case it would necessitate a different, and likely more involved, proof technique than the one used here. An alternative approach would also be to work with the dual reformulation of the DRO problem given in Theorem 2 or to use an alternative probability distance measure rather than f-divergence balls that may be more amenable to this type of analysis.

These approaches come with their own complications, however, and likely require answering similar questions. We believe these are interesting questions worthy of further research, but their investigation is beyond the scope of this work. Instead, we present empirical work which demonstrates the convergence of RDRO in a variety of scenarios.

Chapter 4

Experiments

In this chapter we present several experiments comparing ERM and DRO as well as RRM and RDRO. RRM and RDRO refer to repeatedly optimizing an ERM or DRO objective, respectively, and are defined in Sections 2.4 and 3.1. For simplicity, we will abuse our terminology and avoid the use of RRM and RDRO. Instead, we use ERM to apply to both empirical risk minimization in the supervised learning scenario, as well as at repeated risk minimization (RRM) in the performative prediction scenario. Similarly, we will use DRO to refer to distributionally robust optimization in the supervised learning scenario, as well as repeated distributionally robust optimization (RDRO) in the performative prediction scenario. It will be clear from context which scenario we are discussing.

Our experiments are intended as a proof of concept for the use of DRO in a performative prediction setting, rather than an attempt to demonstrate state-of-the-art performance on a particular task. To this end, we begin in Section 4.2 with an experiment from Perdomo *et al.* [3] which satisfies the assumptions of γ-strong convexity, β-joint smoothness, and ϵ-sensitivity and examine the convergence of ERM and DRO under these conditions. This first experiment is intended as an empirical confirmation of the theory developed in Sections 3.1 and 3.2.

We then move away from strict adherence to these assumptions and explore the convergence behaviour of ERM and DRO with distribution maps that are not neces-

sarily ϵ-sensitive. In Section 4.3 we use simple synthetic datasets to provide intuition for the differing behaviour of ERM and DRO in both the regression and classification settings. The simple regression experiments allow us to build some understanding of important differences between ERM and DRO in a performative context. We restrict the classification experiment in this section to the supervised learning setting to demonstrate properties of ERM and DRO with implications for fairness before introducing the additional complexity of performative prediction.

Finally, in Section 4.4 we explore an example designed to reflect scenarios in which fairness concerns are relevant in order to analyze the behaviour of ERM and DRO from a fairness perspective. We examine ERM and DRO in the performative prediction setting with data composed of majority and minority subgroups. While the data used is still relatively simple, this experiment is intended to capture some of the essential characteristics of machine learning applications in which fairness concerns are relevant, as the potential for discrimination against certain demographic groups is high.

4.1 Implementation

We provide a brief explanation of the implementation details for ERM and DRO before discussing our experiments and their results. Both risk minimization and distributionally robust optimization require access to the data generating distribution, *i.e.* infinite data, which is obviously impossible, so we perform empirical risk minimization and utilize the plug-in estimator for distributionally robust optimization instead. Further, for DRO we make use of the dual formulation (Theorem 2) and perform optimization on this objective rather than working with the primal form. Following Duchi and Namkoong [15] and Hashimoto *et al.* [40], we use χ^2-divergence balls in our implementation, although investigating different choices of f-divergence balls is an interesting future direction.

In all of our experiments we use linear models trained with gradient descent. For

the small datasets we use backtracking line search [41] to adaptively select a step-size and for the larger datasets we use a fixed step-size and a fixed number of epochs for training. The linear regression experiments are ordinary least squares regression, and the logistic regression utilizes an L2-regularized cross-entropy loss function.

The loss functions for an individual data point are specified below. The regression experiments are simply a mean prediction task, so we use x to represent the value of the data and θ to represent the predicted value. The loss for the regression models is:

$$\ell(X;\theta) = (\theta - x)^2.$$

The classification experiments involve trying to predict the true label, $Y \in \{0, 1\}$, from the given covariates, $X \in \mathbb{R}^n$, for each data point. Our data can thus be represented by the random variables $Z = (X, Y)$, and $\hat{y}$ represents the predicted value from a logistic regression model. The loss function is:

$$\ell(Z;\theta) = -y \log(\hat{y}) - (1-y)\log(1-\hat{y}) + \frac{\lambda}{2}||\theta||_2^2$$

The dual formulation of the distributionally robust objective allows for a simple training procedure where we treat the dual variable η as a hyperparameter. Recall that for convex losses, $\ell(\theta; Z)$, the dual formulation is jointly convex in (θ, η). The training procedure is thus as follows, for a given η, compute the approximate mnimimizer $\widehat{\theta_\eta}$

$$\underset{\theta \in \Theta}{\text{minimize}}\, \mathbb{E}(\ell(\theta; Z) - \eta)_+^2.$$

Because the dual is convex in η we can use binary search to find the optimal dual variable η^*. The loss $(\ell(\theta; Z) - \eta)_+^2$ is merely the ReLu function applied to the usual loss with η subtracted and then squared which allows us to train models using gradient descent methods. Hence, using binary search we train models with different values of η until we find the optimal η^*. The optimal value, η^*, depends on the data and also

the radius of our f-divergence ball, ρ. The value of ρ is a hyperparameter that we must specify before training our model.

Selecting the value of ρ is a challenging decision in the implementation of DRO. Different values of ρ will result in very different models and it is difficult to know *a priori* which value of ρ should be selected. There are some theoretical considerations that can be used to inform the decision, but these do not necessarily provide the best guidance. Duchi and Namkoong [15] discuss the choice of ρ and give some practical suggestions for how to select the hyperparameter. Different choices of ρ can result in models whose performance varies significantly, both on the data as a whole, and on distinct subsets of the data.

4.2 Convergence of ERM and DRO for Credit Dataset

Our first experiment reproduces the experimental work from Perdomo *et al.* [3]. We use a dataset from a Kaggle competition titled *Give Me Some Credit*. The dataset contains relevant information for predicting credit scores. The target variable is a binary variable indicating whether or not an individual has experienced financial distress in the past two years. The original dataset contains historical information for 250,000 borrowers.

Following Perdomo *et al.* [3] we balance the dataset to contain an equal number of positive and negative cases for the target variable and normalize predictor variables to have a mean of zero and variance of one. The reduced dataset contains 18,358 entries.

We train L2-regularized logistic regression models on the data for both ERM and DRO. Both models are trained with stochastic gradient descent with a fixed step-size of $\alpha = 0.03$ for 5000 epochs. The step-size and number of epochs were chosen empirically to approximately replicate performance from Perdomo *et al.* [3] on the base distribution. Additionally, a fixed value of ρ was chosen as a radius of the χ^2-divergence ball for DRO. The value of ρ was chosen so that the accuracy of the DRO

model on the full dataset was significantly, but not drastically, different than that of the ERM model.

In order to add a performative element to the prediction problem, we identify 3 of the 10 features as strategic features which will be altered as a function of the parameters of our model. Strategic classification is a particular instance of performative prediction in which individuals adversarially adjust their features to maximize the likelihood of classification to the positive class. Strategic classification has been explored in relation to fairness concerns in previous work [36–39, 42, 43].

As described in Perdomo *et al.* and Hardt [3, 36], we assume individuals have linear utilities $u(\theta, x) = -\langle \theta, x \rangle$ and quadratic costs $c(x', x) = \frac{1}{2\epsilon}||x' - x||_2^2$. The constant ϵ controls the cost individuals incur by altering their features. Individuals thus pay a cost to manipulate their features in order to minimize the likelihood of the model predicting that they will default on their loan, but are unable to change the true outcome, $y \in \{0, 1\}$, of whether or not they default. Given linear utilities and quadratic costs as described here, the individuals' best response is to manipulate their features as

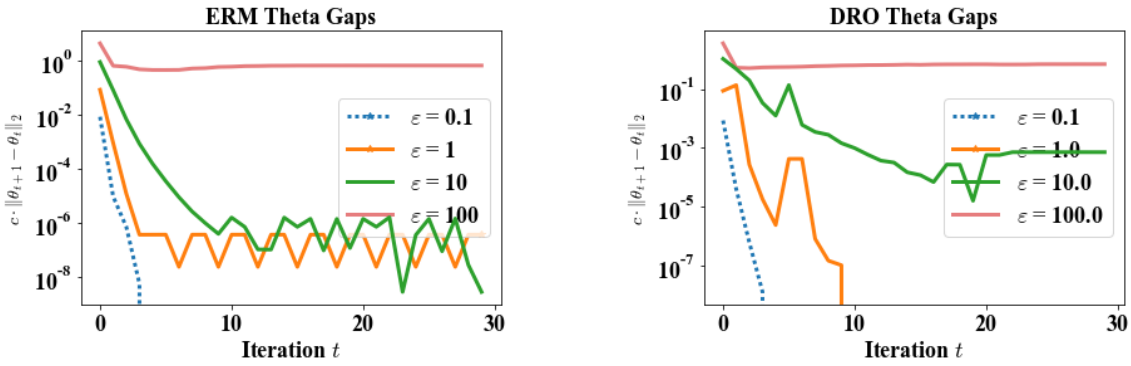
$$x'_S = x_S - \epsilon\theta_S,$$

where x_S, x'_S $\mathbb{R}^{|S|}$ and $|S|$ is the number of strategic features. The explanation of why this distribution map is ϵ-sensitive can be found in Perdomo *et al.* [3].

The procedure for updating the data according to the distribution map for strategic classification is explained in the box below.

Input: Base distribution P, a classifier f_θ, a cost function c, a utility function u.
Sampling procedure for $\mathcal{D}(\theta)$:

1. Sample $(x, y) \sim P$
2. Compute best response $x_{BR} \leftarrow \arg\max_{x'} u(x', \theta) - c(x', x)$
3. Output sample (x_{BR}, y)



(a) ERM (b) DRO

Figure 4.1: Plots of the normalized distance between successive values of θ for ERM and DRO.

Using a logistic regression classifier and the strategic classification distribution map sampling procedure outlined above, we run ERM and DRO for 30 iterations on our dataset with values of $\epsilon \in \{0.1, 1, 10, 100\}$. We observe similar convergence behaviour for both ERM and DRO. As the value of ϵ grows, the inequality $\epsilon < \frac{\gamma}{\beta}$ no longer holds, meaning that the conditions of Theorems 9 and 21 are not satisfied and we do not necessarily have a contraction mapping. We plot the normalized distance between values of θ for successive iterations of ERM and DRO in Figures 4.1a and 4.1b. The distance between iterates is calculated as

$$\frac{1}{||\theta_S||_2} \cdot ||\theta_{t+1} - \theta_t||_2,$$

where θ_S is the value of θ_0 on the strategic features.

In Figures 4.1a and 4.1b we observe that both ERM and DRO converge for $\epsilon = 0.1$ and fail to converge for $\epsilon \in \{10, 100\}$ over 30 iterations of repeated training and deployment. Interestingly, DRO converges for $\epsilon = 1$ while ERM does not appear to converge, but instead appears to cycle between values of θ very close to one another. It is important to note, however, that this behaviour could be due to numerical issues as the gaps between successive iterates are on the order of 10^{-7}. Also, it is worth noting that because we have a fixed step-size and fixed number of epochs, we are not performing exact empirical risk minimization, nor exact distributionally robust

optimization, although the approximate solutions should be very close to the optimal solutions.

This experiment suggests that although we are unable to confirm the assumptions required for the proof of convergence of DRO, it appears that DRO behaves similarly to ERM in terms of convergence. The average accuracy for DRO was approximately 5 percentage points lower than for ERM (approximately 67% vs 73%) throughout the training and testing iterations for the models which converged to performatively stable points. The *Give Me Some Credit* dataset does not contain any demographic information, so we are unable to make any conclusions regarding the fairness properties of either ERM or DRO, but we explore performance on subgroups within data in the upcoming Sections (4.3 and 4.4).

4.3 Building Intuition Through Simple Examples

We now investigate the convergence of ERM and DRO on simple synthetic datasets to build intuition for the behaviour of the two algorithms. We move away from the assumptions required for our general theoretical guarantees for convergence to a fixed point and experiment with distribution maps that are not necessarily ϵ-sensitive. We begin with a regression problem and then investigate a classification problem.

4.3.1 Regression

We start with a simple mean prediction task with data sampled from a mixture of two univariate Gaussians, $X_A \sim \mathcal{N}(\mu_A, \sigma^2)$ and $X_B \sim \mathcal{N}(\mu_B, \sigma^2)$. 80% of the data is sampled from X_A and 20% from X_B, *i.e.* $X = 0.2X_A + 0.8X_B$. We train a linear regression model using gradient descent with backtracking line search to predict the mean of the data. We initialize our data with values $\mu_A = 4$, $\mu_B = 4$, and $\sigma^2 = 0.01$. We choose a small value of σ^2 so that the variance and finite samples have only a small impact on the convergence behaviour of ERM and DRO. We select an f-divergence ball radius of $\rho = 4$ for DRO. The value of ρ is somewhat inconsequential for this

experiment, but we provide further discussion of values of ρ in section 4.4.

We use θ to denote the learned parameter of the model, μ to denote the true mean of the data generating distribution, and μ_A and μ_B to denote the true means of the X_A and X_B. In other words, $\theta = \hat{\mu}$ is the estimated mean from the model. Where necessary, we indicate with subscripts ERM and DRO to indicate which model we are referring to.

We investigate three different distribution maps which we call $\mathcal{D}_0$, $\mathcal{D}_1$, and $\mathcal{D}_2$. For each map the means of the normal distributions from which we sample are adjusted as a function of θ. Hence, the induced distribution is

$$\mathcal{D}_i(\theta) = 0.2\mathcal{N}(\mathcal{D}_i^A(\theta), \sigma^2) + 0.8\mathcal{N}(\mathcal{D}_i^B(\theta), \sigma^2).$$

These distribution maps were chosen because they are simple to understand, but reveal important differences in the way ERM and DRO behave when their predictions alter the distribution on which they are learning. The distribution maps are specified in Table 4.1.

$\mathcal{D}_0(\cdot)$	$D_1(\cdot)$	$\mathcal{D}_2(\cdot)$
$\mathcal{D}_0^A(\theta) = \mathcal{N}(\theta, \sigma^2)$	$\mathcal{D}_1^A(\theta) = \mathcal{N}(\mu_A, \sigma^2)$	$\mathcal{D}_2^A(\theta) = \mathcal{N}(2\theta, \sigma^2)$
$\mathcal{D}_0^B(\theta) = \mathcal{N}(\frac{\theta}{2}, \sigma^2)$	$\mathcal{D}_1^B(\theta) = \mathcal{N}(\frac{\theta}{2}, \sigma^2)$	$\mathcal{D}_2^B(\theta) = \mathcal{N}(\frac{\theta}{2}, \sigma^2)$

Table 4.1: Distribution maps for mean-prediction experiment.

The distribution map determines the evolution of the distribution over the data at each iteration of deploying and learning our model, which in turn changes the learned $\theta_t = \hat{\mu}_t$. Despite starting from the same initial distribution in each case, the induced distributions vary widely. This demonstrates the importance of performativity and illustrates why a static supervised learning framework fails to adequately capture the complex dynamics of prediction problems which involve performativity.

We run the experiment for 50 iterations, where one iteration involves training a model on data sampled from the current distribution and then updating the distribu-

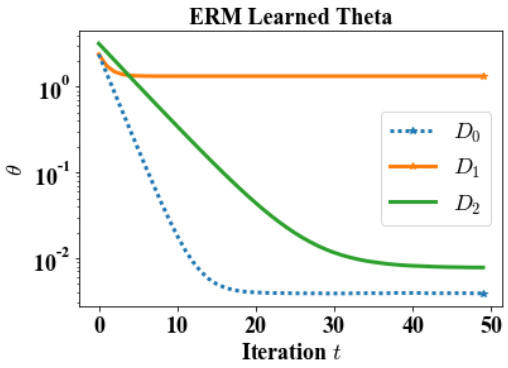


Figure 4.2: Learned values of θ for ERM.

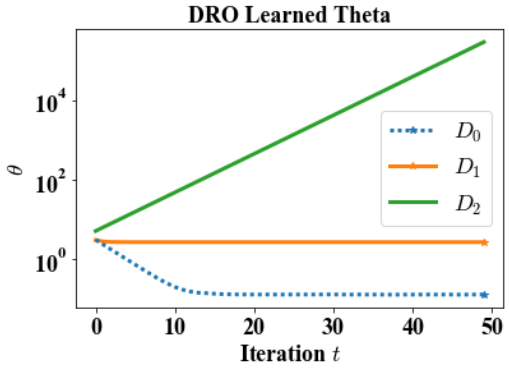


Figure 4.3: Learned values of θ for DRO.

tion via the distribution map. We summarize the evolution of the learned parameter θ_t over time for both ERM and DRO in Figures 4.2 and 4.3. Given the simplicity of the learning problem, the learned values of θ closely approximate true mean of the distribution, μ, over time. The approximate values to which ERM and DRO converge are given in Table 4.2.

It is interesting to note that even with this simple mean prediction task we observe significant differences in the behaviour of ERM and DRO over time. For all three distribution maps we observe ERM's tendency to focus on the majority group over the minority group. Recall that 20% of the data is sampled from X_A and 80% from X_B so we refer to group A as the minority group and group B as the majority group. For $\mathcal{D}_0$, ERM quickly converges toward zero as group B's mean evolves as $\mu_B = \frac{\theta}{2}$. Interestingly, however, it does not converge to zero, but instead remains fixed at approximately $\theta = 0.004$ even though $\theta = 0$ is the performatively optimal value.

DRO displays similar behaviour for this distribution map, but does not converge as close to zero. As noted in Section 2.4, performative stability and performative optimality are distinct solution concepts, and performatively optimal points only coincide with performatively stable points in specific settings.

The second distribution map, $\mathcal{D}_1$, reveals differences in ERM and DRO that are relevant for fairness. For $\mathcal{D}_1$, the mean of the majority group again evolves as $\mu_B = \frac{\theta}{2}$,

Method	$\mathcal{D}_0(\cdot)$	$D_1(\cdot)$	$\mathcal{D}_2(\cdot)$
ERM	$\theta_{ERM} = 0.004$	$\theta_{ERM} = 1.33$	$\theta_{ERM} = 0.008$
DRO	$\theta_{DRO} = 0.128$	$\theta_{DRO} = 2.66$	$\theta_{DRO} = \infty$

Table 4.2: Values to which θ converges for ERM and DRO.

but this time the minority group's mean remains unchanged with $\mu_A = 4$. As ERM averages the loss over all data points, it converges to a predicted mean much closer to μ_B than to μ_A. DRO, on the other hand, converges to a value that balances performance on prediction of the means of both minority and majority groups. In fact the distance from the true means of group A and group B is almost identical.

$$|\theta_{DRO} - \mu_A| = |2.66 - 4| = 1.34$$

$$|\theta_{DRO} - \mu_B| = |2.66 - 1.33| = 1.33$$

ERM, however, is a much better predictor of the global mean of the data. This simple example illustrates a trade-off we can expect between ERM and DRO in terms of fairness versus global performance.

$$\mu_{ERM} = 0.2\mu_A + 0.8\mu_B = (0.2)(4) + (0.8)(0.665) = 1.332$$

$$\mu_{DRO} = 0.2\mu_A + 0.8\mu_B = (0.2)(4) + (0.8)(0.1.33) = 1.864$$

The final distribution map, $\mathcal{D}_2$, demonstrates that ERM and DRO can behave entirely differently under some circumstances. For this distribution map ERM converges to a similar value to that under $\mathcal{D}_0$, that is $\theta_{ERM} = 0.008$. This makes some sense intuitively as $\mu_A = 2\theta$ whereas for $\mathcal{D}_0$ we had $\mu_A = \theta$. DRO, on the other hand, diverges with θ_{DRO} going to infinity. The reason for this is that for the first iteration DRO learns a value of θ_{DRO} that is larger than 4. Similarly, for each successive iteration the learned value of θ_{DRO} gets larger which in turn causes the true means of the data to grow as a function of the learned θ_{DRO}. Interestingly, if you swap the majority and minority groups the behaviour of DRO remains nearly identical, while

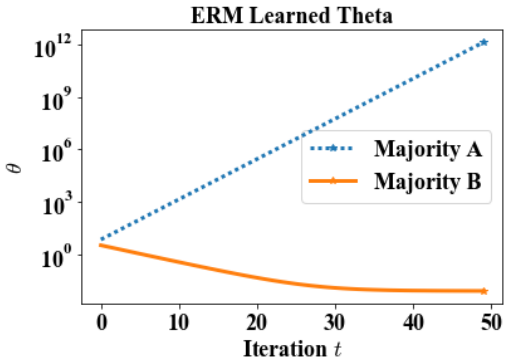


Figure 4.4: Learned values of θ for ERM with $\mathcal{D}_2$.

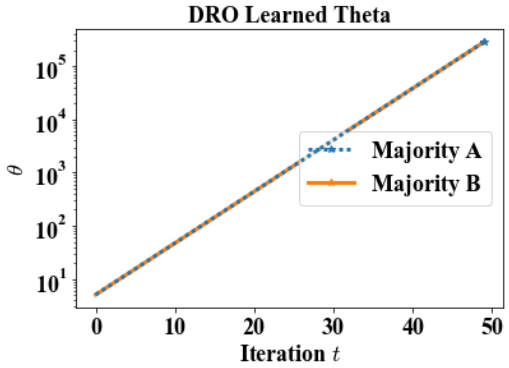


Figure 4.5: Learned values of θ for DRO with $\mathcal{D}_2$.

ERM diverges at a faster rate than DRO. This again demonstrates how DRO treats minority and majority groups similarly, while ERM learns a function that prioritizes performance on majority groups. We illustrate this in Figures 4.4 and 4.5

4.3.2 Classification

We now move to a classification task and analyze the behaviour of logistic regression classifiers trained using ERM and DRO. The classification task is more complex than the simple mean prediction task, so for this experiment we analyze only the static supervised learning setting in order to reduce complexity and elucidate the relevant differences between ERM and DRO. Section 4.4 explores a classification task in the performative prediction setting.

Our data is generated from bivariate Gaussian distributions and the label, y, of a given data point is 1 if the sum of its features are greater than the sum of the means of the two Gaussian distributions from which we draw samples and 0 otherwise. As with the regression experiments, we have two subgroups within our data, A and B. We vary the proportion of samples from each subgroup in the experiments. Precisely,

the data generating process is as follows:

$$X_A \sim \mathcal{N}(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A)$$
$$X_B \sim \mathcal{N}(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$$
$$X = c_A X_A + c_B X_B \quad c_A, c_B \in (0,1), \text{ and } c_A + c_B = 1$$

where

$$\boldsymbol{\mu}_A = \begin{bmatrix} \mu_A^1 \\ \mu_A^2 \end{bmatrix}, \qquad \boldsymbol{\mu}_B = \begin{bmatrix} \mu_B^1 \\ \mu_B^2 \end{bmatrix}, \qquad \boldsymbol{\Sigma}_A = \begin{bmatrix} \sigma_A^1 & 0 \\ 0 & \sigma_A^2 \end{bmatrix}, \qquad \boldsymbol{\Sigma}_B = \begin{bmatrix} \sigma_B^1 & 0 \\ 0 & \sigma_B^2 \end{bmatrix}.$$

And for a data point, $x = [x_i^1, x_i^2]^T$, with $i \in \{A, B\}$,

$$y = \begin{cases} 0 & \text{if } \; x_i^1 + x_i^2 \leq \mu_i^1 + \mu_i^2 \\ 1 & \text{if } \; x_i^1 + x_i^2 > \mu_i^1 + \mu_i^2 \end{cases}.$$

Hence, if $\boldsymbol{\mu}_A \neq \boldsymbol{\mu}_B$, the data is not linearly separable and the logistic regression model must trade off performance across the two subgroups. We provide scatter plots of samples from the data generating distribution below with $\mu_A^i = 1$ and $\mu_B^i = 0.7$, $\sigma_A^i = \sigma_B^i = 0.1$ for $i \in \{0, 1\}$, and $c_A = 0.8$, $c_B = 0.2$. Figure 4.6 contains a sample of 360 data points coloured by the value of their target variable, with crosses representing data points belonging to group B and circles representing data points belonging to group A. Figure 4.7 contains a sample of 500 and 100 data points respectively from group A and group B coloured by the value of their target variable.

In Figure 4.6 we can see that the data is not linearly separable, as some members of group B who belong to the positive class have features that place them lower than the threshold for positive classification for group A. Although this dataset is extremely simple, it is characterized by a feature that represents a central concern for fairness in machine learning, namely that the conditional probability distributions, $P(y|x)$, are significantly different for distinct subsets of the data. This example is intended to represent a simplified abstract instance of a population with majority and minority subgroups in order to see how the behaviour of ERM and DRO differ in this circumstance.

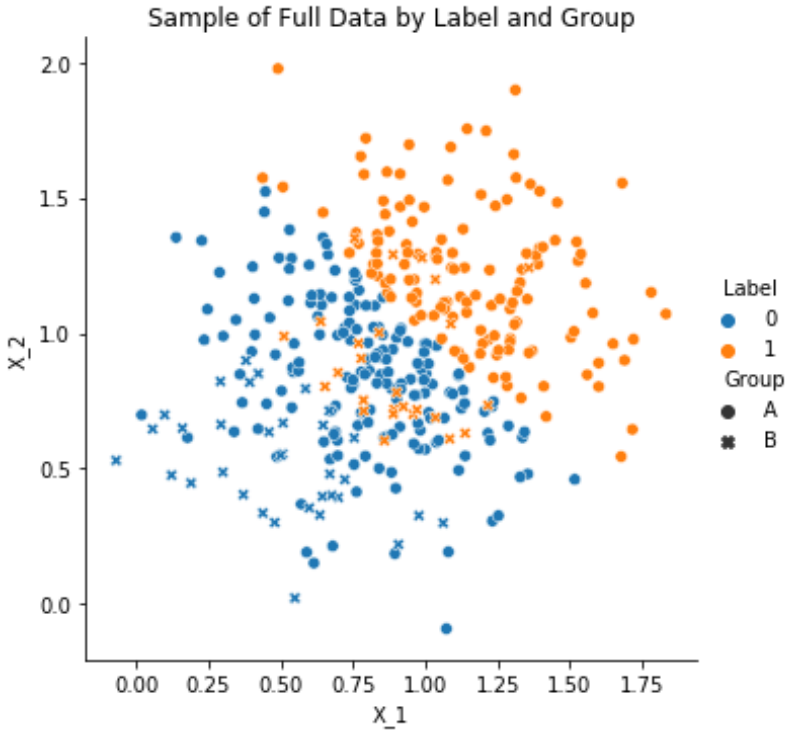


Figure 4.6: Sample of 360 data points from the data generating distribution with $c_A = 0.8$ and $c_B = 0.2$.

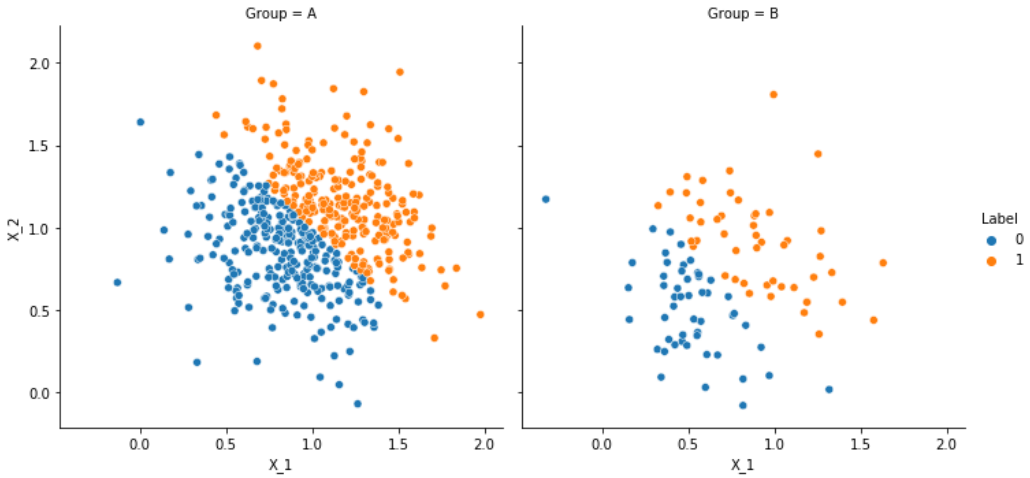


Figure 4.7: Sample of 600 data points from the data generating distribution with $c_A = 0.8$ and $c_B = 0.2$.

We generate three distinct datasets on which to train our algorithms, each made up of differing proportions of the two subgroups A and B. Each dataset contains a sample of 10,000 data points, with samples distributed according the values of c_A and c_B. The accuracy of the models on the three datasets is summarized in Tables 4.3 and 4.4 below. For both ERM and DRO we use L2-regularized logistic regression trained with stochastic gradient descent. The step-size for all algorithms is fixed at 0.05 and we train for 15,000 epochs.

Models trained with the distributionally robust objective have the additional complication that we must specify a value for the radius of the χ^2-divergence ball, *i.e.* ρ. The larger the value of ρ, the more we can expect a DRO model to differ from an ERM model because as the χ^2-divergence ball grows, the worst case distribution can be further and further from the data generating distribution. Conversely, in the limit as $\rho \to 0$, we recover ERM as the χ^2-divergence ball shrinks to contain only the data generating distribution.

Choosing the value of ρ is a challenging decision, as the performance of a model varies significantly as ρ changes. If one has access to demographic information, it is possible to conduct a grid search over possible ρ values in order to find a value that results in a model with the desirable fairness properties. Doing this, however, largely defeats the purpose of DRO. As explained earlier, a central advantage to using DRO rather than some fairness constrained optimization technique is that DRO does not require access to demographic information. In this experiment we work directly with the dual formulation of DRO and set $\eta = 0.56$. This value was chosen empirically to achieve relatively uniform accuracy across group A and group B for an 80/20 split between the two subgroups. As the values of c_A and c_B change, we can see that the performance of DRO changes for a given value of η and hence ρ, as η^* depends on ρ.

We first examine the performance of ERM (Table 4.3). As the data is not linearly separable, ERM must learn a decision boundary that trades off performance between the two subgroups. Because the ERM objective treats the loss on each data point

Group	$[c_A = 0.6,\ c_B = 0.4]$	$[c_A = 0.8,\ c_B = 0.2]$	$[c_A = 0.95,\ c_B = 0.05]$
A	0.797	0.907	0.966
B	0.701	0.652	0.592
All Data	0.759	0.856	0.948

Table 4.3: Accuracy by Group for ERM.

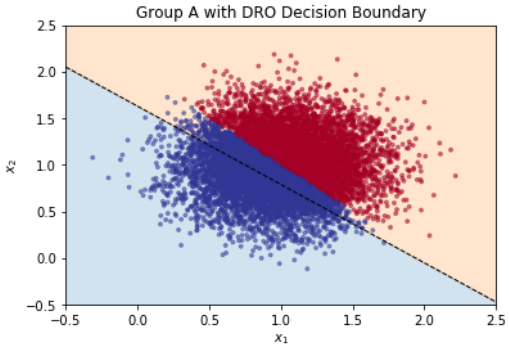
Group	$[c_A = 0.6,\ c_B = 0.4]$	$[c_A = 0.8,\ c_B = 0.2]$	$[c_A = 0.95,\ c_B = 0.05]$
A	0.665	0.751	0.780
B	0.869	0.744	0.766
All Data	0.747	0.750	0.780

Table 4.4: Accuracy by Group for DRO.

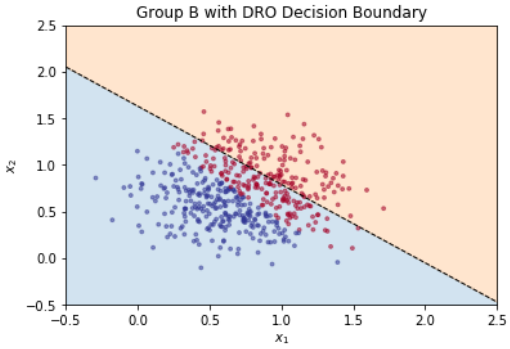
equally, the model learns a decision boundary that is more accurate for the majority group than for the minority group. This discrepancy in accuracy of predictions worsens the smaller the majority group is. For instance, when 95% of the data comes from group A, the logistic regression model trained with an ERM objective achieves 96% accuracy on group A members, but only 59.2% accuracy on group B members.

Models trained with a DRO objective behave much differently. For the 80/20 split and 95/5 split, the DRO models learn relatively fair decision boundaries, effectively balancing performance on both subgroups A and B. For the 60/40 split, however, DRO actually learns a model that performs significantly better on the minority group than the majority group. This model is in a sense discriminatory against the majority group. This is the result of the radius of the χ^2-divergence ball being too large and the model thus overly focusing on a worst case distribution.

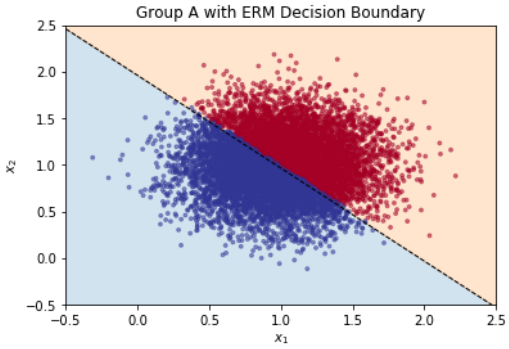
As mentioned above, the correct choice of ρ is not necessarily obvious, but it greatly impacts the performance of the model. Duchi *et al.* provide some recommendations and heuristics for choosing values of ρ in Duchi and Namkoong [15]. Along with



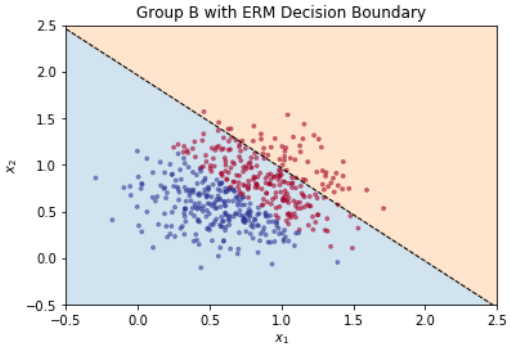
(a) Group A, DRO.



(b) Group B, DRO.



(c) Group A, ERM.



(d) Group B, ERM.

Figure 4.8: Decision boundaries for ERM and DRO classifiers with samples from groups A and B with $c_A = 0.95$ and $c_B = 0.05$. Shading indicates predicted label and data point colour indicates true label.

altering the value of ρ, one can change the f-divergence ball by varying the the value of k for f-divergences in the Cressie-Read family. This has a similar effect to changing the value of ρ.

In Figures 4.8a, 4.8b, 4.8c, and 4.8d we plot data from groups A and B for data generated with $c_A = 0.95$, $c_B = 0.05$ with the learned decision boundary from a DRO and ERM model, respectively, overlaid. The background shading indicates the predicted label with blue representing $\hat{y} = 0$ and red representing $\hat{y} = 1$, while the colour of the data points indicate their true label.

Neither the regression nor the classification experiments in this section are designed to be realistic representations of real-world applications, but rather are intended to provide a simple setting in which to investigate important differences in the behaviour

of models trained with ERM versus DRO objectives. In the next section we investigate a slightly more complex classification task in the performative prediction setting and examine how DRO and ERM may impact fairness considerations when deploying a model influences the distribution on which it is making predictions.

4.4 Fairness and DRO vs ERM

We now examine ERM and DRO from a fairness perspective. We generate synthetic data in the same manner as the previous static classification experiment, except with 10 dimensional multivariate Gaussians as opposed to 2 dimensions. Unlike the classification experiment in Section 4.3, however, we examine a performative prediction task rather than a static classification task. To be precise, the data generating process is as follows:

$$
\begin{aligned}
X_A &\sim \mathcal{N}(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A) \\
X_B &\sim \mathcal{N}(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B) \\
X &= c_A X_A + c_B X_B \quad c_A, c_B \in (0,1), \text{ and } c_A + c_B = 1
\end{aligned}
$$

where

$$
\boldsymbol{\mu}_A = \begin{bmatrix} \mu_A^1 \\ \vdots \\ \mu_A^{10} \end{bmatrix}, \quad \boldsymbol{\mu}_B = \begin{bmatrix} \mu_B^1 \\ \vdots \\ \mu_B^{10} \end{bmatrix}, \quad \boldsymbol{\Sigma}_A = \begin{bmatrix} \sigma_A^1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_A^{10} \end{bmatrix}, \quad \boldsymbol{\Sigma}_B = \begin{bmatrix} \sigma_B^1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_B^{10} \end{bmatrix}.
$$

And for a data point, $x = [x_i^1, \ldots, x_i^{10}]^T$, with $i \in \{A, B\}$,

$$
y = \begin{cases} 0 & \text{if } \; x_i^1 + \cdots + x_i^{10} \leq \mu_i^1 + \cdots + \mu_i^{10} \\ 1 & \text{if } \; x_i^1 + \cdots + x_i^{10} > \mu_i^1 + \cdots + \mu_i^{10} \end{cases}.
$$

For our experiment we set the parameters of the data generating process as

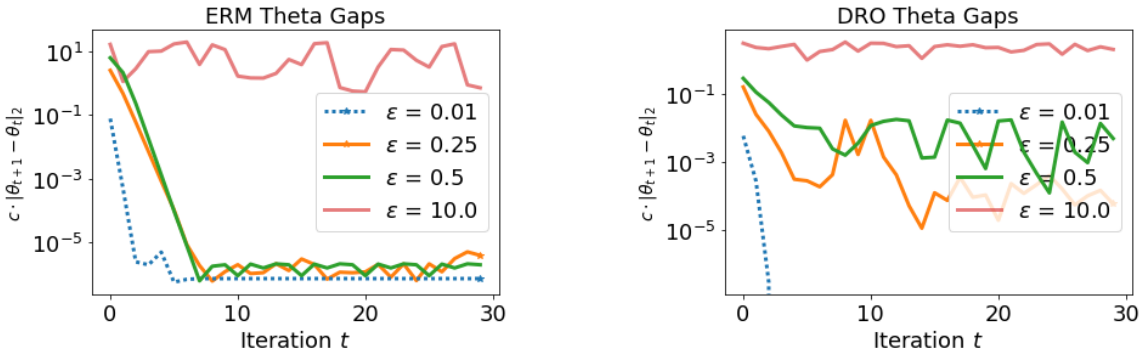
$$
\begin{aligned}
&\mu_A^i = 1 && \sigma_A^i = 0.1 && c_A = 0.8 \\
&\mu_B^i = 0.8, && \sigma_B^i = 0.1, && c_B = 0.2.
\end{aligned}
$$

These parameters were selected as an attempt to capture the notion that an underprivileged minority group may have features that make them appear to be less qualified, despite having the same proportion of qualified individuals as a dominant majority group. The data generating process obviously represents this at a high level of abstraction and is much less complex than most real-world applications. With that said, we believe the data generating process effectively encapsulates this abstracted characteristic that is central to concerns for learning fair models. Examples of the type of situation that this experiment is intended to represent are college admissions or hiring, where an underprivileged minority group may not, on average, have CVs that appear as impressive as their peers from the majority group due to a lack of opportunity, but are nevertheless equally qualified for the school or job.

We once again examine a strategic classification scenario, as in Section 4.2, but our data now contains subgroups, allowing us to analyze the impact of ERM and DRO on fairness and model performance in the performative setting. The explanation of the distribution map induced by strategic classification can be found in Section 4.2. We examine four different distribution maps by varying the parameter $\epsilon \in \{0.01, 0.25, 0.5, 10\}$ and set 5 of the 10 features to be strategic (*i.e.* manipulable).

For both ERM and DRO we train L2-regularized logistic regression models with $\lambda = 0.0001$. We use stochastic gradient descent with a fixed step-size of $\alpha = 0.2$ and train for 8000 epochs on samples of 1,200 data points. We use a fixed radius ρ of the χ^2-divergence ball for DRO. All parameters were chosen empirically to give good performance on the base distribution.

First, we examine the convergence behaviour of both algorithms by plotting the normalized distance between successive iterates of the learned parameters, θ_t, in Figures 4.9a and 4.9b. We observe that ERM does not converge for any value of ϵ, while DRO converges for only $\epsilon = 0.01$. It is unclear why the algorithms do not converge for other values of ϵ. The failure to converge could be related to using a fixed, rather than decaying step-size, or because the conditions for the contraction mapping are not



(a) ERM (b) DRO

Figure 4.9: Plots of the normalized distance between successive values of θ for ERM and DRO.

met (recall that the conditions are sufficient, but not necessary for convergence of a particular instance). For the three smaller values of ϵ, both ERM and DRO converge to a small neighbourhood, whereas for $\epsilon = 10$, neither ERM nor DRO exhibit any convergence.

To demonstrate the effect of the convergence, or lack thereof, of θ on the model's performance, we plot the average supervised and performative L2-regularized binary cross-entropy loss for ERM and DRO for $\epsilon = 0.5$ and $\epsilon = 10$ in Figure 4.10. The blue lines indicate the optimization phase and the green lines indicate the effect of the distribution shift after the classifier deployment. That is, the dots at the end of a green dotted line represent performative loss, while the dots at the end of a blue line represent supervised loss. In the plots we can see that even though ERM and DRO do not converge for $\epsilon = 0.5$, the models quickly achieve relatively stable loss on the classification task. Note that the average L2-regularized binary cross-entropy loss is the objective for which ERM is optimizing, but not the objective for which DRO is optimizing.

Given that this experiment is a balanced binary classification task, the metric in which we are principally interested is accuracy. Furthermore, as we are interested in comparing the fairness properties of DRO and ERM, it is important to analyze the performance of the models on the subgroups within the population, as well as the

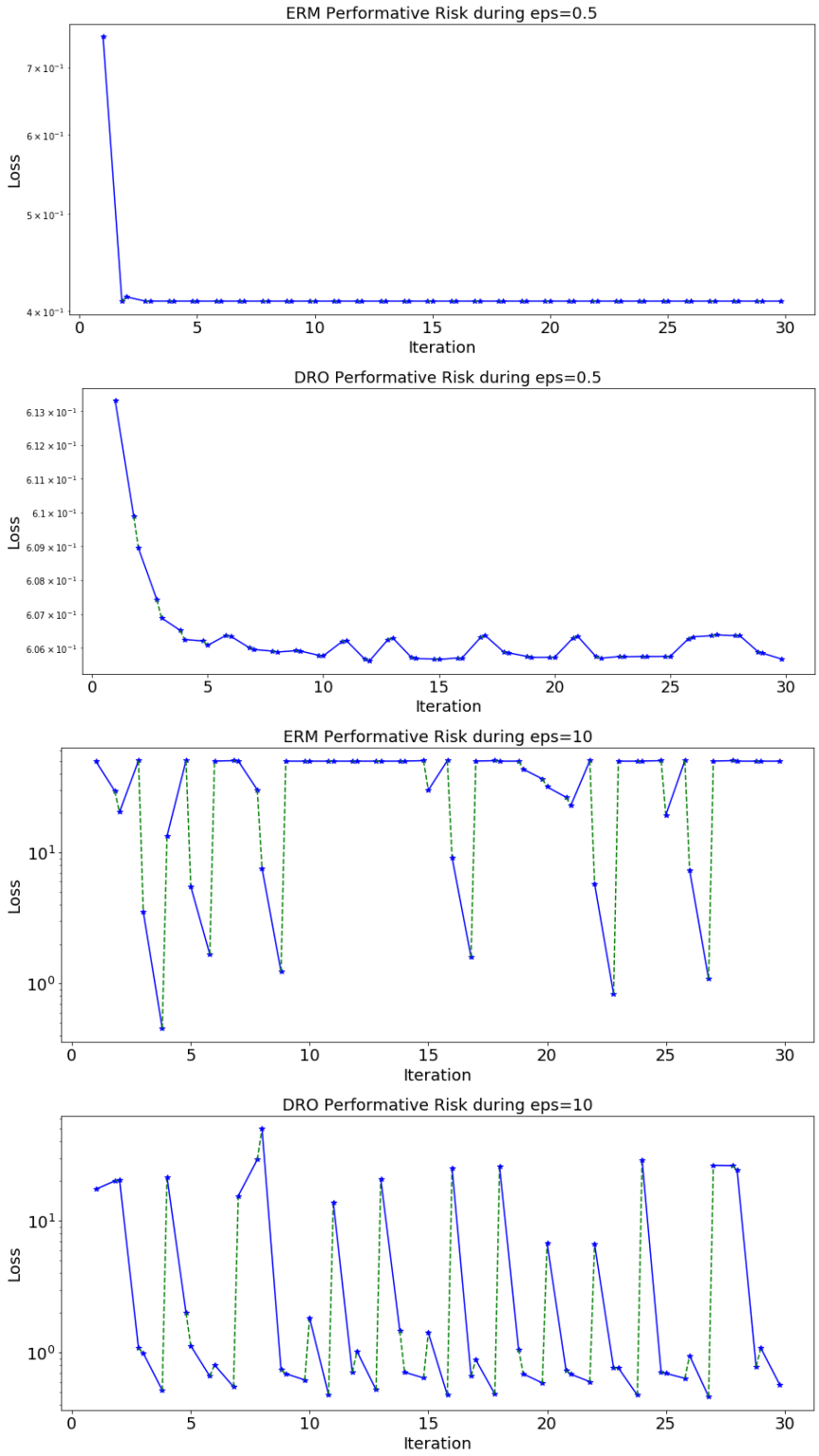


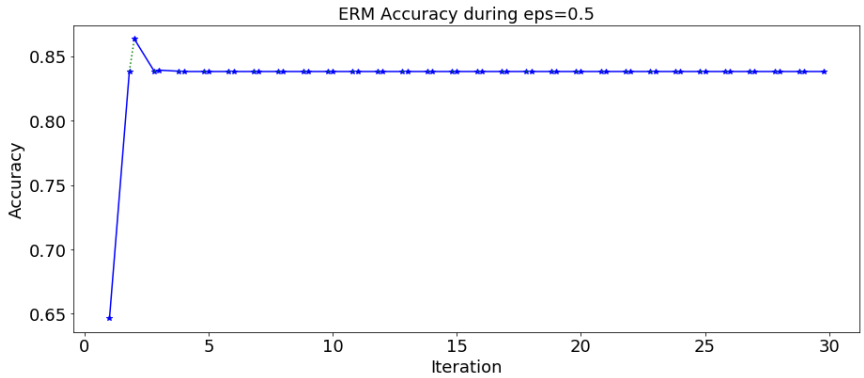
Figure 4.10: Plots of the average L2-regularized binary cross-entropy supervised and performative loss.

population as a whole. In Figures 4.11a and 4.11b we plot the the accuracy of the two models for $\epsilon = 0.5$. We see that the performative accuracy of ERM initially degrades significantly, before quickly converging to approximately 84% on the full population.

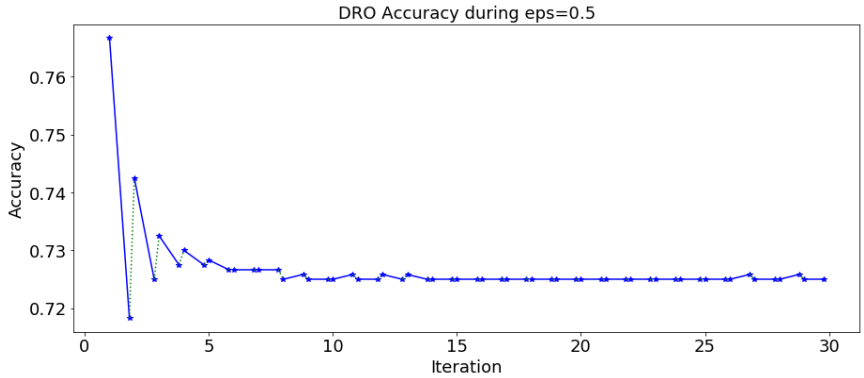
Conversely, the performative accuracy for DRO is actually initially higher than the accuracy on the distribution on which the model was trained, but it also converges relatively quickly to an accuracy of approximately 72.5% on the full population. This is once again due to DRO having a different optimization objective than ERM. The average cross-entropy loss acts as a surrogate loss for the 0-1 loss and hence aims to maximize accuracy on the full dataset. The objective used for DRO focuses on the tails of the distribution and therefore does not maximize accuracy across the full dataset. The improvement of the DRO model's accuracy after the distribution shift is not an inherent property of DRO, but rather is specific to this particular dataset and distribution map. We see in Figure 4.11a at iteration $t = 2$ that ERM also occasionally achieves better performance after distribution shift. In general, the improvement or degradation of accuracy depends on the learned parameters of the model, the distribution map, the data, and the loss function.

We saw in the static classification experiment in Section 4.3 that ERM achieved higher global accuracy than DRO on data composed of minority and majority subgroups, but did so because it focused on learning a good model for the majority group at the expense of the minority group. DRO, on the other hand, balanced performance on both subgroups. The natural question to ask is if these dynamics carry over to the performative prediction setting with higher dimensional data. In Figure 4.12 we plot the accuracy of ERM and DRO for the subgroups A and B for $\epsilon = 0.5$. We can see from these plots that the fairness properties of ERM and DRO are preserved in the performative prediction setting.

ERM converges to an accuracy of approximately 90% on group A and only approximately 54% on group B. This discrepancy is even worse than what we saw in the 2-dimensional static classification experiment in Section 4.3. DRO, on the other



(a) ERM



(b) DRO

Figure 4.11: ERM and DRO accuracy on the full population across successive iterations.

hand, converges to much more equal accuracy across the two subgroups, achieving an accuracy of approximately 74% for group A and approximately 66% for group B. We summarize the accuracies to which ERM and DRO converge in Tables 4.5 and 4.6. We do not include the accuracy for $\epsilon = 10$ because neither ERM nor DRO converged to a small enough neighbourhood, and therefore did not converge in accuracy. We see that the algorithms converge to similar accuracy values for all the values of ϵ.

This result, while perhaps not that surprising, is extremely important, as it demonstrates that not only does DRO exhibit similar convergence behaviour to ERM, but DRO converges to fair fixed points, whereas ERM converges to discriminatory fixed points in the presence of heterogeneous data composed of minority and majority subgroups. Recall also that DRO is not given access to group information, but still

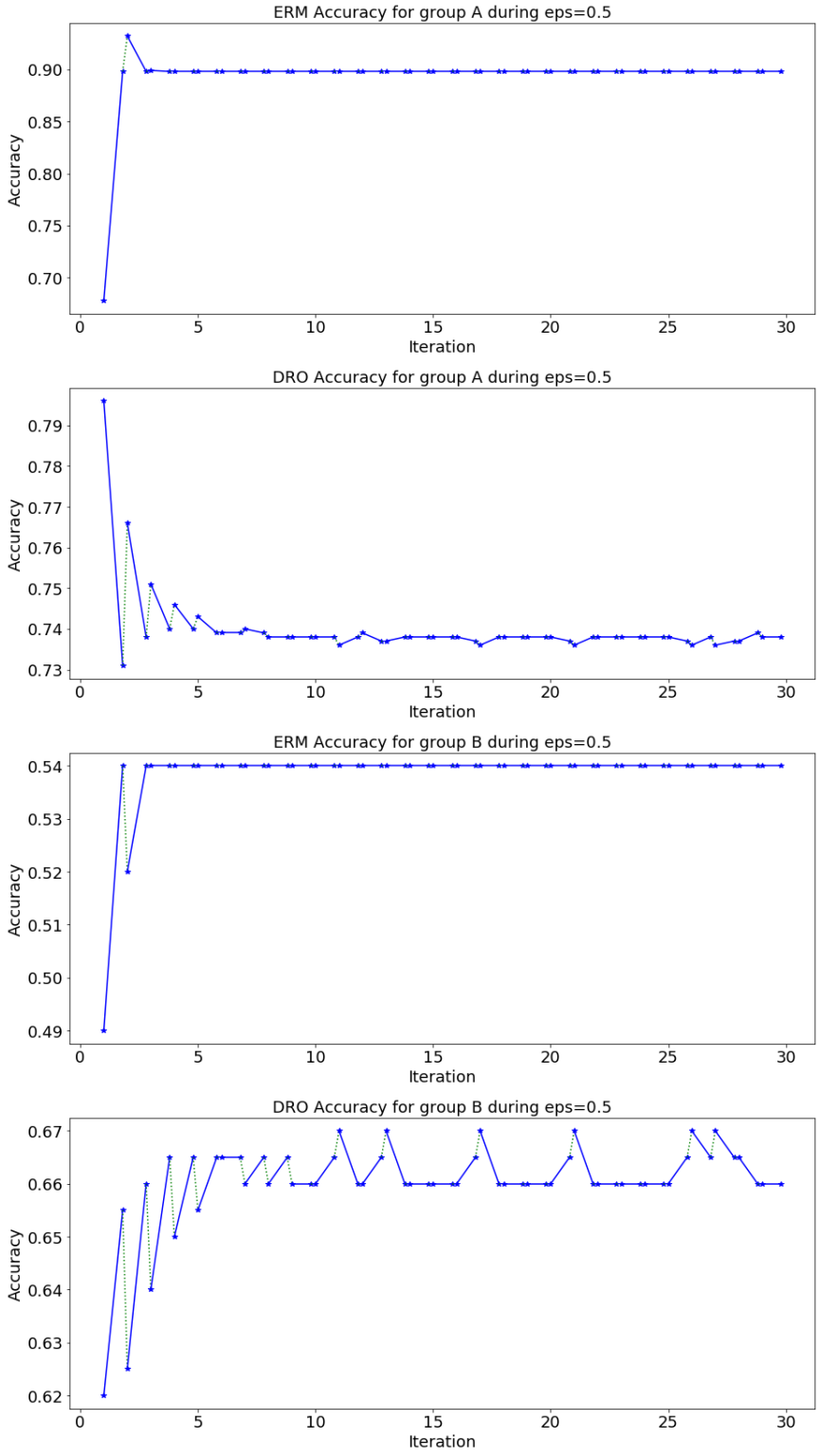


Figure 4.12: ERM and DRO accuracy on subgroups A and B across successive iterations.

learns to achieve more uniform performance across subgroups, as it is attempting to minimize the worst case loss across all probability distributions within the χ^2-diverge ball surrounding the data generating distribution.

ERM Performative Accuracy

Group	$\epsilon = 0.01$	$\epsilon = 0.25$	$\epsilon = 0.5$
A	0.893	0.896	0.898
B	0.540	0.540	0.540
All Data	0.834	0.837	0.838

Table 4.5: Accuracy by Group for ERM after 30 iterations.

DRO Performative Accuracy

Group	$\epsilon = 0.01$	$\epsilon = 0.25$	$\epsilon = 0.5$
A	0.687	0.710	0.738
B	0.670	0.660	0.660
All Data	0.684	0.701	0.725

Table 4.6: Accuracy by Group for DRO after 30 iterations.

As we discussed at the beginning of this thesis in Section 2.1, fairness in machine learning is a contested notion, with competing formal definitions of what it means for a machine learning algorithm to be fair. We have avoided explicitly weighing in on this debate in this research, but in arguing that DRO results in less discriminatory or more fair models than ERM we are implicitly endorsing a particular notion of fairness. We believe, however, that fairness fundamentally depends on context and that there is thus no universal definition of fairness.

This experiment is intended as an abstraction of a fairness-relevant machine learning scenario and we make several underlying assumptions about what the data is intended to represent. In this example, we assume that the labels for the data points

are in fact the correct labels. That is, if a data point has a label $y = 1$, that data point should indeed be labelled in the positive class. This assumption allows us to reasonably argue that a "fair" classifier is one which achieves uniform performance across subgroups. Of course, there is more subtlety to the issue. For instance, the model learned by the algorithm with a DRO objective results in more false positives for data points in group A and more false negatives for data points in group B because the classifier learns a decision boundary in between the hyperplanes which separate positive and negative classes for group A and group B.

Given that we feel none of the current formal fairness criteria are adequate for capturing the complex notion of fairness in machine learning, we are not concerned that our notion of "fairness" in this work does not necessarily adhere to any of these criteria. We instead intend to make a more general case for the fairness properties of DRO as compared to ERM when dealing with heterogeneous data composed of majority and minority subgroups.

Additional plots for the experiment in this section can be found in Appendix A.

Chapter 5

Conclusion

5.1 Discussion

This thesis has explored, and attempted to contribute to, three new and emerging areas of research in machine learning: fairness, distributionally robust optimization, and performative prediction. Fairness research in machine learning has thus far largely focused on developing formal fairness definitions for static supervised classification problems. In recent years, however, there has been a realization of the limitations of this work and attempts to understand fairness beyond the static supervised learning scenario, as well as to address some of the shortcomings of the mainstream formal fairness criteria.

Performative prediction [3] represents an attempt to lay a theoretical foundation and create a framework for fairness researchers to build on. Early work in the performative prediction space has exclusively explored optimization objectives that minimize average loss, such as empirical risk minimization. As we discuss in this work, techniques that focus on average loss are susceptible to learning discriminatory models when the data generating process is composed of minority and majority subgroups characterized by different conditional probability distributions. In this thesis we build on this research by replacing ERM with a distributionally robust objective [15].

We extend definitions from Perdomo *et al.* [3] to show that under analogous assumptions to those proposed there, repeated distributionally robust optimization ex-

hibits similar convergence behaviour to that of repeated risk minimization. We then empirically verify this result, before moving away from strict adherence to the assumptions required for Theorems 9 and 21. Finally, we design an experiment to demonstrate that using a distributionally robust objective has the potential to train fair machine learning models without access to demographic information and over the long-term in the presence of changing distributions.

The experiments in Chapter 4 are not intended as a comprehensive demonstration that DRO is superior to ERM for achieving fairer models when making predictions about heterogeneous data, rather their goal is a proof of concept of DRO as method for achieving fairness in a dynamic world characterized by changing distributions. If DRO failed to converge to performatively stable points under any circumstance, or was unable to learn models that successfully balance uniform performance across subgroups with overall model performance on even simple datasets it would be an indication that DRO is not a good candidate for designing fair algorithms. Our experiments, however, showed that DRO does indeed converge to performatively stable fixed points and that models trained with DRO objectives can very successfully learn models that achieve relatively uniform performance across subgroups, even in the presence of changing distributions. There is much more work to be done to further explore these ideas, but we have attempted to demonstrate that this is indeed a promising line of research with the potential to resolve some of the current issues of fairness research in machine learning.

In Section 2.1.1, we discussed important shortcomings of the current formal fairness criteria which we briefly recall here:

1. It is unclear which definition should be used.

2. The definitions apply to static supervised classification problems only.

3. We often do not have access to demographic information.

4. The definitions ignore intersectionality.

We have attempted to address these shortcomings in this work and we hope that it acts as a building block towards a more complete notion of fairness in machine learning and the development of techniques to ensure we are not training discriminatory algorithms.

5.2 Future Work

This work represents a preliminary investigation into the combination of DRO and performative prediction as a method to understand and learn fair algorithms, and, as such, leaves many important avenues for future work open.

On the theoretical side, there is much more work to be done in understanding the convergence properties of DRO. The result in this work eschewed answering important questions of stability of worst case distributions as well as questions of smoothness and strong convexity of distributionally robust objectives. There are also a number of other results in the performative prediction literature that have been shown for risk minimization but not for other optimization objectives. Attempting to extend these results to DRO is an interesting area of future work.

Another interesting direction is the exploration of alternative robust objectives other than DRO. DRO represents only one of many robust optimization techniques and it could well be that other robust objectives are more appropriate for the performative prediction framework and for addressing fairness concerns in machine learning.

Similarly, it is not necessarily clear that performative prediction is the correct framework for understanding fairness in machine learning. While it offers a much richer and more complex setting than static supervised learning, performative prediction is still limited in that distribution maps must be functions of the parameters of a model, θ, rather than the loss incurred by a model, $\ell(\theta, Z)$. Many scenarios where one would expect to encounter undesirable feedback loops from deploying ML mod-

els are those in which the model's performance on the data directly effects the data distribution. This effectively adds a notion of "state" to the performative prediction problem which adds a significant degree of complexity. This complexity complicates the mathematics, but also allows for richer models of real-world scenarios.

There is also much more work to be done on the empirical side of things. The experiments in this thesis are intended as a proof of concept rather than a demonstration on real-world problems. Many areas where fairness is a concern involve complex data and complex moral, legal, and philosophical questions regarding what fairness means in that particular context. Examining the effectiveness of DRO in these scenarios, and developing methodology for selecting important hyperparameters such as the type of f-divergence to use and choosing a value of ρ is an important area of research if the use of DRO is to be adopted by ML practitioners.

Finally, performative prediction and distributionally robust optimization have connections and similarities to many other well established areas of research in machine learning, and it is likely that existing results in areas such as bandits, reinforcement learning, online learning, stochastic optimization, and out of distribution generalization could greatly contribute to our understanding of DRO and performative prediction.

5.3 Final Thoughts

We began this work with a discussion of why research into fairness in machine learning is such an urgent issue and we would like to reemphasize this point. Machine learning is no longer a niche research area. It has become a fundamental driving force of the global economy and now attracts billions of dollars in investment around the world [44, 45]. This change has resulted in enormous opportunity for individuals with expertise in machine learning and it appears that this trend will only increase in the coming years.

The reality, however, is that this emergence of AI as an economic powerhouse has

not occurred in a particularly responsible manner. Machine learning researchers and engineers have benefited greatly from the investment in AI and the community has a responsibility to ensure that the future development of the field aligns with and supports universal rights, freedoms, and values. Chief among these responsibilities is ensuring that the development and deployment of machine learning models do not harm the most vulnerable among us. We hope that this work contributes to this endeavor.

Bibliography

[1] J. Horowitz *et al.*, "The facebook files," *Wall Street Journal*, 2021. [Online]. Available: https://www.wsj.com/articles/the-facebook-files-11631713039.

[2] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2019, http://www.fairmlbook.org.

[3] J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt, "Performative prediction," in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., ser. Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, pp. 7599–7609.

[4] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5 2, pp. 153–163, 2017.

[5] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," ser. KDD '17, Association for Computing Machinery, 2017, 797–806, ISBN: 9781450348874. DOI: 10.1145/3097983.3098095.

[6] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ser. ITCS '12, Association for Computing Machinery, 2012, 214–226, ISBN: 9781450311151. DOI: 10.1145/2090236.2090255.

[7] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *NIPS*, 2016.

[8] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociological Methods & Research*, vol. 50, no. 1, pp. 3–44, 2021. DOI: 10.1177/0049124118782533.

[9] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th International Conference on World Wide Web*, ser. WWW '17, International World Wide Web Conferences Steering Committee, 2017, 1171–1180, ISBN: 9781450349130. DOI: 10.1145/3038912.3052660.

[10] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, C. H. Papadimitriou, Ed., ser. Leibniz International Proceedings in Informatics (LIPIcs), vol. 67, Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017, 43:1–43:23, ISBN: 978-3-95977-029-3. DOI: 10.4230/LIPIcs.ITCS.2017.43.

[11] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro, "Learning non-discriminatory predictors," in *Proceedings of the 2017 Conference on Learning Theory*, S. Kale and O. Shamir, Eds., ser. Proceedings of Machine Learning Research, vol. 65, PMLR, 2017, pp. 1920–1953.

[12] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999. DOI: 10.1017/CBO9780511624216.

[13] V. Vapnik, "Principles of risk minimization for learning theory," in *Proceedings of the 4th International Conference on Neural Information Processing Systems*, ser. NIPS'91, Morgan Kaufmann Publishers Inc., 1991, 831–838, ISBN: 1558602224.

[14] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *FAT*, 2018.

[15] J. C. Duchi and H. Namkoong, "Learning models with uniform performance via distributionally robust optimization," *The Annals of Statistics*, vol. 49, no. 3, pp. 1378 –1406, 2021. DOI: 10.1214/20-AOS2004.

[16] A. Wald, "Statistical decision functions which minimize the maximum risk," *Annals of Mathematics*, vol. 46, pp. 265–280, 1945.

[17] D. Wozabal, "A framework for optimization under ambiguity," *Annals of Operations Research*, vol. 193, no. 1, pp. 21–47, 2012.

[18] G. Pflug and D. Wozabal, "Ambiguity in portfolio selection," *Quantitative Finance*, vol. 7, no. 4, pp. 435–442, 2007.

[19] J. Lee and M. Raginsky, "Minimax statistical learning with wasserstein distances," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc., 2018.

[20] A. Renyi, "On measures of entropy and information," in *The 4th Berkeley Symposium on Mathematics, Statistics and Probability*, University of California Press, 1961, 547–561.

[21] I. Csiszár and P. Shields, "Information theory and statistics: A tutorial," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 4, pp. 417–528, 2004, ISSN: 1567-2190. DOI: 10.1561/0100000004.

[22] I. Csiszar, "Information-type measures of difference of probability distributions and indirect observation," *Studia Scientifica Mathematica*, vol. 2, 299–318, 1967.

[23] A. Shapiro, "Distributionally robust stochastic programming," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2258–2275, 2017. DOI: 10.1137/16M1058297.

[24] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, Mar. 2004, ISBN: 0521833787.

[25] J. Rawls, *A Theory of Justice*, 1st ed. Belknap Press of Harvard University Press, 1971, ISBN: 0-674-88014-5.

[26] C. Kroer, *Lecture Notes for Economics, AI, and Optimization*. 2022.

[27] C. Mendler-Dünner, J. Perdomo, T. Zrnic, and M. Hardt, "Stochastic optimization for performative prediction," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 4929–4939.

[28] J. Miller, J. C. Perdomo, and T. Zrnic, "Outside the echo chamber: Optimizing the performative risk," *International Conference on Machine Learning (ICML) 2021*,

[29] G. Brown, S. Hod, and I. Kalemaj, "Performative prediction in a stateful world," 2020. arXiv: 2011.03885.

[30] R. Dong and L. J. Ratliff, *Approximate regions of attraction in learning with decision-dependent distributions*, 2021. arXiv: 2107.00055 [cs.LG].

[31] H. Namkoong and J. C. Duchi, "Variance-based regularization with convex objectives," in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/5a142a55461d5fef016acfb927fee0bd-Paper.pdf.

[32] J. C. Duchi, T. Hashimoto, and H. Namkoong, "Distributionally robust losses for latent covariate mixtures," *CoRR*, vol. abs/2007.13982, 2020. arXiv: 2007.13982. [Online]. Available: https://arxiv.org/abs/2007.13982.

[33] S. Jeong and H. Namkoong, "Robust causal inference under covariate shift via worst-case subpopulation treatment effects," in *Proceedings of Thirty Third Conference on Learning Theory*, J. Abernethy and S. Agarwal, Eds., ser. Proceedings of Machine Learning Research, vol. 125, PMLR, 2020, pp. 2079–2084. [Online]. Available: https://proceedings.mlr.press/v125/jeong20a.html.

[34] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust Optimization*, ser. Princeton Series in Applied Mathematics. Princeton University Press, 2009.

[35] A. Ben-Tal, D. den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen, "Robust solutions of optimization problems affected by uncertain probabilities," *Management Science*, vol. 59, no. 2, 341–357, 2013, ISSN: 0025-1909. DOI: 10.1287/mnsc.1120.1641. [Online]. Available: https://doi.org/10.1287/mnsc.1120.1641.

[36] M. Hardt, N. Megiddo, C. Papadimitriou, and M. Wootters, "Strategic classification," in *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ser. ITCS '16, Cambridge, Massachusetts, USA: Association for Computing Machinery, 2016, 111–122, ISBN: 9781450340571. DOI: 10.1145/2840728.2840730. [Online]. Available: https://doi.org/10.1145/2840728.2840730.

[37] S. Milli, J. Miller, A. D. Dragan, and M. Hardt, "The social cost of strategic classification," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19, Association for Computing Machinery, 2019, 230–239, ISBN: 9781450361255. DOI: 10.1145/3287560.3287576. [Online]. Available: https://doi.org/10.1145/3287560.3287576.

[38] L. Hu, N. Immorlica, and J. W. Vaughan, "The disparate effects of strategic manipulation," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19, New York, NY, USA: Association for Computing Machinery, 2019, 259–268, ISBN: 9781450361255. DOI: 10.1145/3287560.3287597. [Online]. Available: https://doi.org/10.1145/3287560.3287597.

[39] A. D'Amour, H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, and Y. Halpern, "Fairness is not static: Deeper understanding of long term fairness via simulation studies," ser. FAT* '20, New York, NY, USA: Association for Computing Machinery, 2020, 525–534, ISBN: 9781450369367. DOI: 10.1145/3351095.3372878. [Online]. Available: https://doi.org/10.1145/3351095.3372878.

[40] T. B. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, "Fairness without demographics in repeated loss minimization," in *ICML*, 2018.

[41] L. Armijo, "Minimization of functions having Lipschitz continuous first partial derivatives.," *Pacific Journal of Mathematics*, vol. 16, no. 1, pp. 1 –3, 1966. DOI: pjm/1102995080. [Online]. Available: https://doi.org/.

[42] M. Bruckner, C. Kanzow, and T. Scheffer, "Static prediction games for adversarial learning problems," *Journal of Machine Learning Research*, vol. 13, no. 85, pp. 2617–2654, 2012. [Online]. Available: http://jmlr.org/papers/v13/brueckner12a.html.

[43] J. Miller, S. Milli, and M. Hardt, "Strategic classification is causal modeling in disguise," in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., ser. Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, pp. 6917–6926. [Online]. Available: https://proceedings.mlr.press/v119/miller20b.html.

[44] A. Abdallat, "A push for digital transformation: The global ai investment boom," *Forbes Technology Council*, 2021. [Online]. Available: https://www.forbes.com/sites/forbestechcouncil/2021/12/22/a-push-for-digital-transformation-the-global-ai-investment-boom/?sh=4091edc10bf2.

[45] T. Balakrishnan, M. Chui, B. Hall, and N. Henke, "The state of ai in 2020," *Mckinsey Analytics*, 2020. [Online]. Available: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2020.

Appendix A: Additional Figures

Included below are figures for all values of ϵ for the experiment in Section 4.4. We include the plots for the distance between θ_t over successive iterations, performative accuracy on the full population, performative accuracy on subgroup A, performative accuracy on subgroup B, and the average L2-regularized cross-entropy loss on the full population.

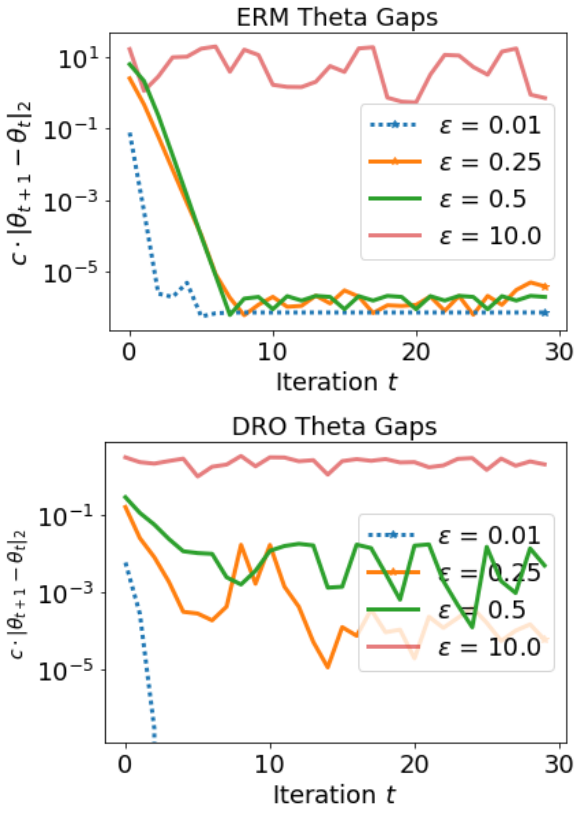


Figure A.1: Plots of the normalized distance between successive values of θ for ERM and DRO.

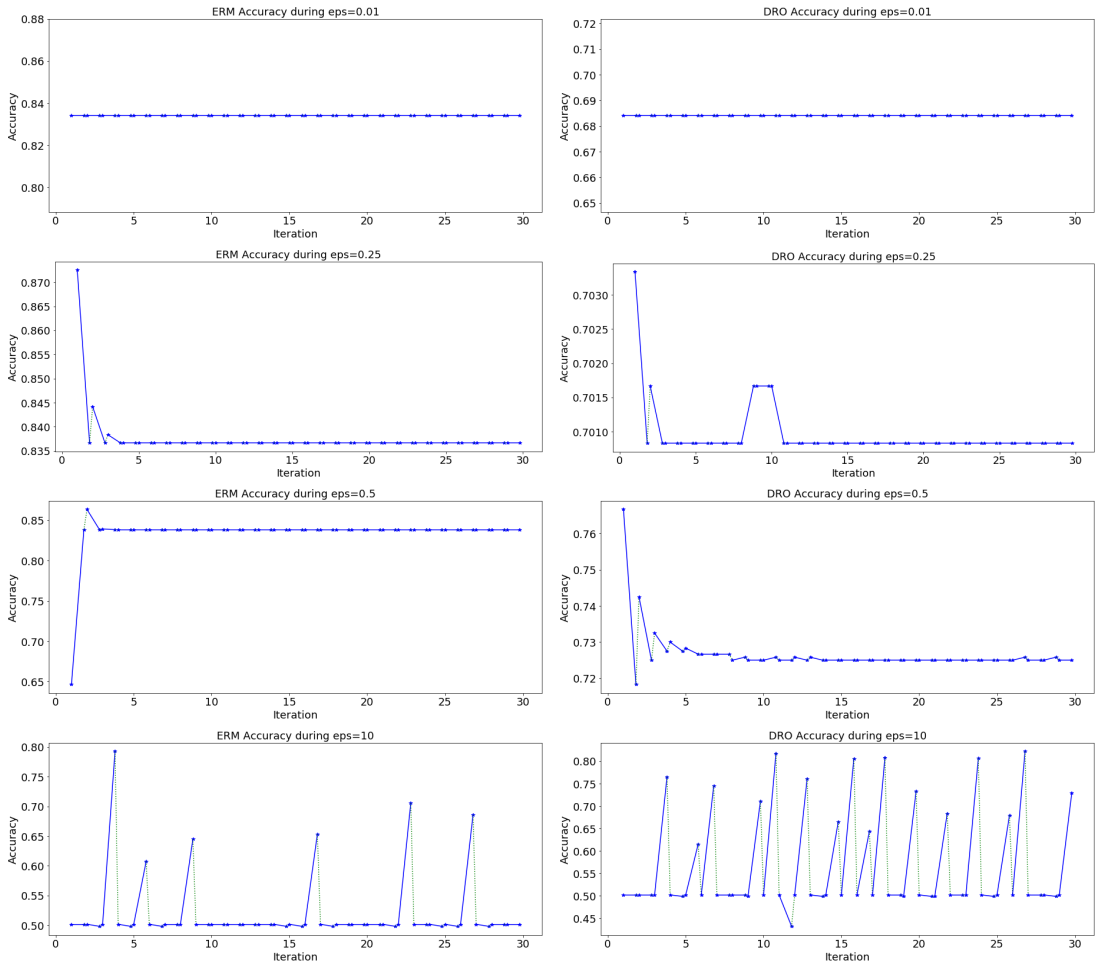


Figure A.2: ERM and DRO accuracy on full population across successive iterations.

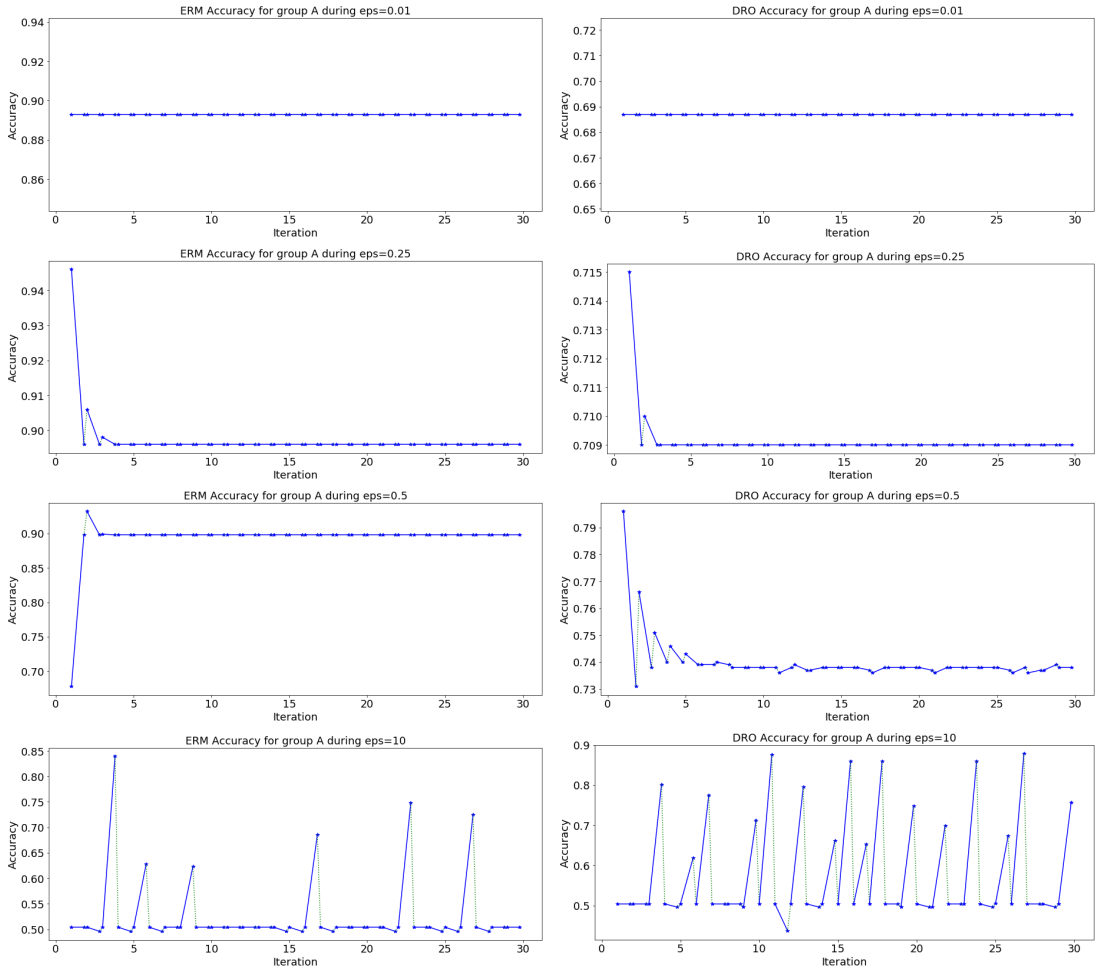


Figure A.3: ERM and DRO accuracy on group A across successive iterations.

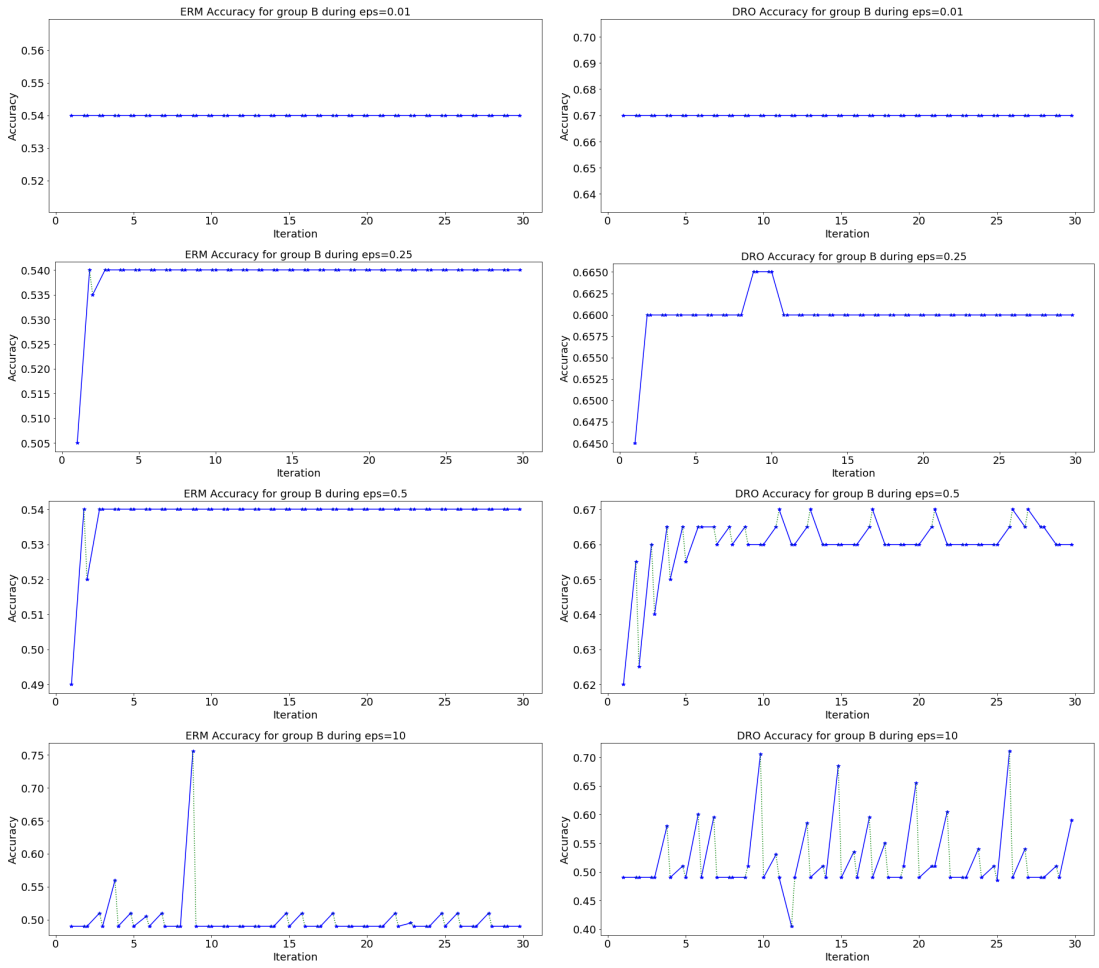


Figure A.4: ERM and DRO accuracy on group B across successive iterations.

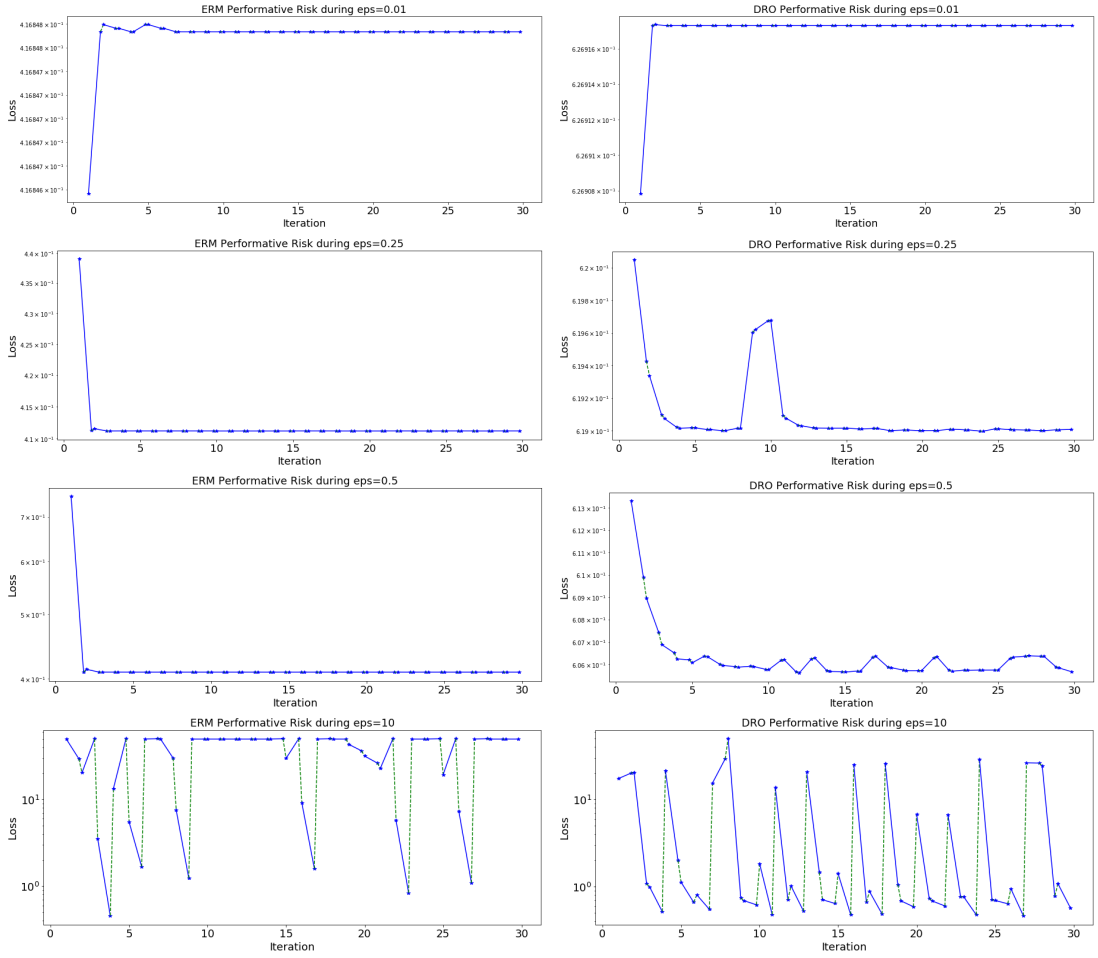


Figure A.5: Plots of the average L2-regularized binary cross-entropy supervised and performative loss for ERM and DRO.