Performative Prediction: Theory and Practice

Juan Perdomo

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2023-80
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2023/EECS-2023-80.html

May 9, 2023

Performative Prediction: Theory and Practice

By

Juan Carlos Perdomo Silva

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering- Electrical Engineering & Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Peter L. Bartlett, Co-chair
Associate Professor Moritz Hardt, Co-chair
Professor Benjamin Recht
Associate Professor Avi Feller

Spring 2023

Performative Prediction: Theory and Practice

Abstract

Performative Prediction: Theory and Practice

by

Juan Carlos Perdomo Silva

Doctor of Philosophy in Engineering- Electrical Engineering & Computer Science

University of California, Berkeley

Professor Peter L. Bartlett, Co-chair

Associate Professor Moritz Hardt, Co-chair

When algorithmic predictions inform social decision-making, these predictions don't just forecast the world around them: they actively shape it. Building models that influence the world is, in fact, often the primary goal of prediction. For example, in medicine, we predict the risk of a person developing a disease to hopefully minimize the likelihood that it occurs. In elections, we predict voting preferences with the goal of targeting information campaigns that are explicitly designed to influence people's political beliefs. If done properly, social predictions are *performative*. They directly interact with the world around them and change it.

In this thesis, we will first introduce a learning-theoretic framework, performative prediction, that places these problems on formal mathematical grounds. We will illustrate how the framework can be used to analyze common social prediction dynamics, such as repeated retraining in response to strategic effects. Furthermore, we will discuss how it can be used to design decision rules that embrace the distinction between forecasting future outcomes accurately and steering them towards socially desirable targets.

In the second part of the thesis, we will use these theoretical ideas as a guiding lens to study early warning systems, a popular class of risk prediction tools used in over half of US public high schools. We will present the results of a collaboration with the Wisconsin Department of Public Instruction in which we performed the first large-scale evaluation of the long-term impacts of early warning systems on graduation rates. At the end, we will close with a discussion regarding the policy implications of our work.

To my parents, Carlos M. Perdomo and María M. Silva

Caminante, son tus huellas
el camino y nada más;
caminante, no hay camino,
se hace camino al andar.

Al andar se hace camino
y al volver la vista atrás;
se ve la senda que nunca
se ha de volver a pisar.

Caminante no hay camino,
sino estelas en la mar.

- Antonio Machado

Contents

Acknowledgments

It somehow feels all too soon to be writing these acknowledgements, my years at Berkeley have gone by so quickly. I can clearly remember my visit days, sitting down for dinner at Angeline's with Chloe and Nilesh, meeting Moritz and Peter, and deciding to start what has truly been a wonderful adventure.

I'm most grateful to my advisors Peter Bartlett and Moritz Hardt for giving me the opportunity to be their student. As someone who spent more time sailing than doing research in undergrad, I'm glad they saw enough in me back then to admit me to grad school. They have been constant sources of wisdom during my years here in Berkeley. I admire Moritz's ability to be such a creative and independent thinker. I remember sitting with him outside of SDH at the end of my first year and telling him I was quite anxious about not having a good research topic yet. He reassured me that we should keep exploring new areas and that inevitably something interesting would turn up. This thesis proves he was right. Peter taught me to appreciate the beauty of tackling fundamental, technical questions. He has been incredibly supportive of me throughout my PhD. I feel very lucky to have had the chance to work together on topics in adaptive control and statistical learning theory, that I would have liked to also include in this thesis.

I'm thankful to have had Ben Recht and Avi Feller on my PhD committee. Their feedback has greatly improved the quality of my work. In addition to teaching me all about optimization in his class, I'm very grateful to Ben for making Soda 5 such a vibrant community and including me as an honorary member of his group. I sincerely appreciate how he was always willing to speak his mind and convey his take on whatever problem I was telling him about. No acknowledgements section would be complete without also thanking Yaron Singer for getting me started in research and having faith in my ability to do interesting work, even when I did not.

The results in this thesis are the consequence of having been able to work with so many amazing collaborators. The theory results in the first part are joint work with Celestine Mendler-Dunner, John Miller, and Tijana Zrnic. They have been excellent collaborators and even better friends. During the pandemic years, I learned so much from working with Max Simchowitz. I am very thankful for his mentorship and patience with me. Then, at the end of grad school, I was very fortunate to have had Michael Kim around Berkeley. He's been somewhat of a third advisor, teaching me how to give better talks and exposing me to new research areas. It's been a pleasure to work together. I am also grateful to have worked with Alekh Agarwal, Akshay Krishnamurthy, and Sham Kakade through the Microsoft Research-BAIR Commons initiative. The second part of this thesis would not have been possible without Rediet Abebe and Tolani Britton. It was one of Rediet's

talks at Berkeley that inspired me to go beyond my theory work and do empirical research at the intersection of machine learning and public policy.

I've been blessed to have such great friends to accompany me during my time in grad school. Nicholas Larus-Stone was there when I took my first CS class in undergrad, and except for a small break when he lived in NYC, has been an amazing roommate and friend. Anibal de la Torre brought that Puerto Rican humor to Berkeley and kept my fridge full of beer during his time here. Nathaniel ver Steeg's sharp wit was always there whenever I took myself too seriously. And, I'm glad Dan Fulop escaped med school and flew out to California so many times to take me backpacking.

Throughout my PhD, I could always count on Nilesh Tripuraneni to be there with some sage advice, or some strong spirits. His infectious laughter was sorely missed throughout Berkeley once he graduated. I am grateful to Aldo Pacchiano, Kush Bhatia, Yeshwanth Cherapanamjeri, and Niladri Chatterji for welcoming me to statlearning and giving me faith that one day, I too could understand what is "regret". I feel lucky to have coincided in grad school with Chloe Hsu, Akosua Busia, Serena Wang, Vickie Ye, Wenshuo Guo, and Clara Wong-Fannjiang. Esther Rolf, Horia Mania, Stephen Tu, Mihaela Curmei, Deb Raji, Paula Gradu, Jessica Dai, Chris Harshaw, Ricardo Sandoval, Anastasios Angelopoulos, Ezinne Nwankwo, Ghassen Jerfel, Frances Ding, Carlos Albors, Karl Krauth, Giacomo Meanti, Lydia Liu, Jiri Hron, Sarah Dean, Smitha Milli, and Yu Sun have been great labmates and friends.

Part of the reason I came to Berkeley initially was to work with Ludwig Schmidt on adversarial robustness. We never did. Instead, we mostly worked on getting faster biking times. However, my views on research have been shaped by the many conversations we've had together and I am thankful for his mentorship.

My algebra teacher at Colegio San Ignacio, Fernando Cruz, would constantly tell us a story that Pythogoras' theorem was not proved by Pythagoras himself, but by one of his students. However, that was "back when students were grateful". So to settle old scores, thank you to Fernando, Iván Flores, and the rest of the CSI math department. I would also like to thank my history teachers Marta Almeida and Jose López for teaching me how to write and think critically.

Coming back to Berkeley after my time at home during the pandemic, the Cal Triathlon team gave me a much needed sense of community. It's been very special to see how Kate Kennedy, Grace Dwyer, and Dean Harper (amongst many others) have maintained such a welcoming environment where people from all backgrounds can enjoy themselves and grow in the sport. All the swim sets at Golden Bear kept me even keeled throughout the ups and downs of the PhD. On that note, I would like to thank Steven Zheng for his friendship and coaching.

I would certainly not have been able to get here without the love and support of my

family. To my mom Marimar, my dad Carlos, my brother Iván, and my sister Carla, I can't thank you enough for believing in me, even if you always didn't exactly know what I was up to. My grandfathers Juan Silva Parra and Diego Perdomo Alvarez would have loved to see me finish my PhD. They are greatly missed. Last (but certainly not least!) I want to thank my girlfriend Lily Katz for being there with me and supporting me throughout all these years. I really feel so lucky.

Chapter 1

Introduction

Whenever algorithmic predictions inform human decision-making, these predictions don't just forecast the world around them: they actively shape it.

Building predictors that influence the world is often the entire point of prediction. In medicine, we predict the risk of a person having a heart attack to hopefully minimize the likelihood that it occurs. In elections, we predict voting preferences with the goal of targeting ads that are explicitly designed to influence people's political beliefs. If done properly, social predictions are *performative*; they directly interact with the world around them and change it.

Once we notice this feedback between predictions and their encompassing environments, we start to see that performative predictions are pervasive. In addition to examples listed above, people use data and algorithmic predictions to make better decisions in finance, education, online job networks, and even in planning for climate change. Anywhere predictions interact with people, changing the behavior of the algorithm directly influences the observable behavior of the broader social system. This observation is the intellectual starting point for the results contained in this thesis.

Despite these rich, dynamic interactions between algorithms and their environments, social prediction problems are typically formalized through the lens of supervised learning. The central assumption in this framework is that the population in question is well approximated by a fixed probability distribution from which individuals and outcomes are drawn randomly. To apply supervised learning in medicine, for example, we must believe that patients and their individual health outcomes are static quantities that we passively observe and make predictions on. Crucially, the predictions we make cannot impact the likelihood of future health outcomes. If we effectively act upon the outputs of our predictor, we invalidate the central tenet of supervised learning!

Neglecting the feedback, or the lack thereof, between algorithmic predictions and their

social environment leads to the design of learning systems that lack validity and fail to address the broader social objectives they were designed for. In the latter part of this thesis, we will illustrate an unfortunate example of these shortcomings in the context of a popular class of risk predictors used within US public education.

We believe that artificial intelligence and machine learning will play important roles in the development of new solutions to challenging social problems. However, the success of this agenda depends on achieving a rigorous understanding of the issues that arise when data-driven algorithms interact with people. Over the last few decades, we've started to see exciting progress in this direction. There has been exciting research formalizing what it means to preserve individual privacy [16], ensure fairness [17], and designing algorithms that are compatible with economic incentives .

In this thesis, we contribute to this broader agenda by studying a different piece of the puzzle that has so far been largely neglected by the academic community: the ways prediction systems can directly influence data distributions. The results of this work are motivated by what we refer to as the performativity thesis of machine learning.

When we apply prediction to social problems, the population in question is not a static object. It is dynamic, and a function of the predictive model.

Performativity is a well-studied concept within the social sciences and has been previously applied to provide new insights in linguistics and economics [6, 51]. In this work, we bring this rich intellectual tradition to learning theory and place performative prediction problems on firm mathematical foundations. As hinted at by the title, the thesis has two main parts: a theoretical component and an empirical case study.

1.1 Part 1: Theoretical Foundations of Performative Prediction

To articulate concerns and propose solutions, we need to be able to talk about problems clearly. Therefore, in the first part of this thesis, we introduce a learning-theoretic framework, performative prediction, which formalizes the idea that predictions can directly change the observed distribution of data.

Sitting in between the generality of reinforcement learning and the simplicity of supervised learning, the framework is carefully scoped to capture the specific types of dynamics present in social prediction problems. On a technical level, we aim to develop a mathematical language that is rich enough to express the impacts of social predictions, yet is not *so broad* that positive results become computationally or statistically intractable.

Performative prediction enables us to make succinct and precise claims regarding the behavior of learning algorithms in social systems. Conceptually, we divide our theoretical insights established through this line of work into two distinct categories. In more detail, in the first part of this thesis we illustrate how the framework enables us to:

1. Analyze the convergence and limiting behavior of common, existing social prediction dynamics, such as *retraining* (Chapter 3).

2. Proactively embrace the causal impacts of prediction to find performatively "optimal" decision rules. Importantly, optimality could entail the desire to *forecast* future outcomes accurately, as well as to *steer* data distributions towards socially desirable targets. (Chapter 4 & 5)

With regards to the first point, when ignored, performativity can surface as a form of distribution shift. As the decision maker acts according to a predictive model, the distribution over data points appears to change over time. For example, in bank lending, people may repeatedly manipulate their features in response to a decision rule with the hopes of achieving a desired outcome. In practice, the response to such distribution shifts is to frequently retrain the predictive model as more data becomes available. Retraining is often considered an undesired — yet necessary — cat and mouse game of chasing a moving target.

Amongst our theoretical contributions, we analyze the closed-loop behavior of this repeated retraining dynamic. We identify conditions under which retraining converges to an equilibrium solution we refer to as *performative stability*, whereby the deployed model is minimizes expected risk for the distribution that it induces. Performatively stable models are fixed points of retraining. Furthermore, under certain conditions they might also have good predictive performance, but not always. Identifying when and why performatively stable models perform well from an accuracy or social welfare perspective is a central focus of our work in Chapter 2.

Moving onto the second point, apart from analyzing existing social prediction dynamics, performative prediction enables the design of new "optimal" learning algorithms that proactively account for the causal impacts of prediction. If predictions can shape the world around us, the design space of "optimal" predictors is significantly more expansive than in supervised learning. In particular, the goals of prediction may not be to just accurately forecast future outcomes, but also to actively steer them towards socially desirable targets.

Drawing upon a recent and exciting line of work on *omniprediction* [28, 29], we demonstrate how one can learn decision rules that are simultaneously "best in class" for many different high-level objectives in performative settings. These learning algorithms effec-

tively serve as a "menu" of *performatively optimal* decision rules that enable decision makers to flexibly decide on the goals of prediction.

Performative optimality is a distinct solution concept from performative stability. Intuitively, it is a wholistic measurement of how well a predictor performs according to some loss function, while considering the fact that different predictors can induce different distributions. Performatively optimal models can, in general, be difficult to solve for, in Chapter 4 we identify several natural conditions under which they can be found efficiently.

Lastly, performative prediction is largely a nascent area of research, and we hope readers will be inspired to establish new connections between this field and other domains. Throughout our presentation, we will aim to mention several exciting works in this direction.

1.2 Part 2: Empirical Investigations

Using these theoretical results on performativity as a guiding lens, in the second part of this thesis we conduct a case study on the use of algorithmic predictions in US public education.

During the past 15 years, there's been a boom in the use of early warning systems to improve low high school graduation rates. Early warning systems are risk assessment tools which predict the probability that each individual student in middle school will graduate from high school on-time. Counselors and teachers used these predictions to identify at-risk students with the goal of individually targeting interventions that keep students on track to graduate.

Using over a decade's worth of data from the DEWS program implemented in Wisconsin public schools, we provide the first large-scale evaluation of the long-term impacts of early warning systems. Quickly summarizing, we find that risk assessments made by the system are highly accurate. Yet, despite being actively examined by schools, we find no evidence that the availability of these predictions has led to improved graduation rates in Wisconsin.

Said otherwise, the risk scores made by the Wisconsin DEWS system are not performative. This is unfortunate since the ideal early warning system *should be* strongly performative. Recall that predictions are explicitly made with the intent of changing educational trajectories for underserved students. Predicting that a student is at high risk of dropping out should ideally lead to an effective intervention that ensures students do graduate on-time. High risk predictions should be self-negating prophecies. This lack of

performativity stems from misunderstanding the social context in which these predictions take place.

In search of an explanation behind the accuracy, but non-impact of the DEWS program, we uncover a robust statistical law present in Wisconsin public schools. Within each school, academic outcomes are essentially independent of individual student performance. Most of the variance in the individual likelihoods of on-time graduation come from the fact that students across schools with different levels of resources and wealth behave very differently. And this variance is indeed captured by environmental features used within DEWS that lead to a high level of overall accuracy.

However, students within the same school have nearly identical features and likelihoods of on-time graduation. The lack of within school variance means that assigning each student in the school the same probability of on-time graduation (i.e. the school average) is a near-optimal prediction. From the perspective of counselors and school staff, the predictions provide, little to no new information and are hence largely ignored. It stands to reason that predictions which are ignored cannot be performative and change outcomes.

These empirical results have direct policy implications for the use of algorithmic systems in education. They demonstrate that, due to the degree of racial and socioeconomic segregation between US school districts, even perfectly implemented individual risk predictors provide little to no actionable information to counselors needing to make difficult decisions regarding which students to devote attention to. The overarching barrier towards improving graduation rates in Wisconsin is not how to identify future dropouts within schools, but rather how to overcome structural differences between schools. We hope that our work inspires future evaluations of algorithmic systems in public policy initiatives.

Part I

Theoretical Foundations of Performative Prediction

Chapter 2

The Performative Prediction Framework

In this chapter, we formally present the performative prediction framework.

Conceptually, the goal of our work here is to introduce a set of mathematical definitions and tools that allow us to clearly reason about the ways predictive models can actively shape their surrounding social environments. In other words, we turn the performativity thesis of machine learning, outlined in the introduction, into formal mathematical terms. The framework brings together different ideas from classical supervised learning, game theory, and optimal control to provide a new perspective on learning in social systems. We will do our best to point out the connections and differences to these different fields as we go along.

The concepts we introduce in this chapter setup the questions we will address later on regarding the long-term behavior of social prediction dynamics (such as retraining) and the design of performatively optimal decision rules. The focus of this chapter, however, is largely not on algorithms, but rather on definitions and illustrating through various examples how these definitions map onto real world concepts we aim to study. We will get to algorithms in later chapters.

In particular, the focus of this chapter will be on discussing the ideas of performative optimality and performative stability. While optimality describes decision rules that are "best in class" according to some user specified notion of optimality, stable solutions are a fixed point definition. They describe the limits of retraining dynamics that often occur in practice when predictions impact the world. While at first glance, these definitions might seem completely unrelated, we will study how they can in fact describe very similar prediction models in certain settings. The main technical results on this chapter describe the similarities and differences between these two definitions with the aim of contextualizing later algorithmic results.

2.1 Main Definitions

Performative prediction is a decision-theoretic framework that extends the classical statistical theory underlying risk minimization. The goal of risk minimization is to find a decision rule, specified by model parameters θ, that performs well on a fixed joint distribution $\mathcal{D}$ over features x and an outcome variable y.

For example, in linear regression where we make predictions $\hat{y} = \theta^\top x$, the goal is to find the vector of parameters $\theta^\star$ that minimizes the average squared error:

$$\theta^\star \in \arg\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[(y - \theta^\top x)^2\right].$$

Throughout our presentation in this chapter, we focus on predictive models f_θ that are parametrized by a vector $\theta \in \Theta$, where the parameter space $\Theta \subseteq \mathbb{R}^d$ is a closed, convex set. In terms of notation, we will refer to predictive models both as θ and f_θ, interchangeably. Lastly, whenever we define a variable $\theta^* = \arg\min_\theta g(\theta)$ as the minimizer of a function g, we resolve the issue of the minimizer not being unique by setting $\theta^\star$ to an arbitrary point in the $\arg\min_\theta g(\theta)$ set.

The Distribution Map

Whenever predictions are performative, the choice of predictive model affects the observed distribution over instances $z = (x, y)$. We formalize this intuitive notion by introducing the idea of a *distribution map* $\mathcal{D}(\cdot)$.

The distribution map is a function from the set of model parameters to the space of distributions. For a given choice of parameters θ, we think of $\mathcal{D}(\theta)$ as the distribution over features and outcomes that results from making decisions according to the model specified by θ. This mapping from predictive model to distribution is the key conceptual device of our framework and mathematizes the performativity thesis outlined in the introduction.

The distribution map is quite powerful in terms of the kinds of dynamics it can describe. In particular, the choice of model can influence the marginal distribution over features X or the conditional distribution over outcomes Y. It can also change both of these quantities simultaneously.

We call a problem *feature performative* if the choice of model f_θ affects the distribution over X.[1] Alternatively, we say a problem is *outcome performative* if the marginal distribution over features is static and unaffected by f_θ, but the conditional distribution over Y is

[1]Formally, performative prediction subsumes *strategic classification* [33]. In this learning problem, an institution makes predictions on people. As a response to the system, individuals manipulate their features with the goal of achieving a desirable classification.

influenced by the predictor. Lastly, we will say a problem is *jointly performative*, or just performative for short, if f_θ influences the joint distribution over (x, y).

Example 2.1.1 (financial trading)**.** To illustrate these different types of performative effects, consider the task of predicting the future price of a commodity (e.g., oil) in order to inform trading decisions. Depending on the choice of features, the problem exhibits different kinds of performativity.

If we use the temperature today to predict the price of oil tomorrow, the features are not performative. However, the future price of oil tomorrow is affected by our prediction since the prediction shapes trading activity which ultimately determines prices. This problem is hence outcome performative, but not feature performative. If in addition to temperature, we also incorporated some index of consumer demand into our feature set, then the problem becomes jointly performative. Demand depends on prices which are in turn influenced by our predictions.

While we might have some intuition regarding what subsets of the observed data are actively influenced by the predictive model, we generally assume that the distribution map $\mathcal{D}(\cdot)$ is *unknown* to the learner. To improve the quality of their model, the learner can however observe the kind of data that it induces. If they deploy a model θ, the learner gets to observe samples $(x, y) \sim \mathcal{D}(\theta)$, where $\mathcal{D}(\theta)$ is some arbitrary distribution that may be significantly different for different models f_θ.

The Performative Risk

Given that different models induce different distributions, a natural objective in performative prediction is to evaluate a model θ on the resulting distribution $\mathcal{D}(\theta)$. Here, performance is measured via some loss function ℓ. This results in a notion we call the *performative risk*, defined as

$$\mathrm{PR}(\theta) := \mathbb{E}_{z\sim\mathcal{D}(\theta)}\ell(z;\theta)\,. \tag{2.1}$$

Relative to objectives like the expected risk is supervised learning, the model parameters θ appear in two places within the expression for the performative risk.[2] This additional dependence of the distribution on the model parameters means that the performative risk may not be convex in θ, even if the loss is a convex function of θ. Understanding when and why the performative risk is convex is the central focus of Chapter 4.

Remark 2.1.2 (regarding loss functions)**.** At a high level, a loss function is a way for us to express preference over (data, prediction) pairs. In performative prediction, the kinds of preferences we want to express may differ significantly from those in supervised learning.

[2] In supervised learning, there is a fixed data distribution $\mathcal{D}$ and the expected risk is: $\mathbb{E}_{z\sim\mathcal{D}}\ell(z,\theta)$.

For example, when performing supervised learning over binary outcomes (assume $\hat{y}, y \in \{0, 1\}$), we often optimize the 0-1 loss $\mathbf{1}\{y \neq \hat{y}\}$ since we prefer *accurate* predictions. That is, predictions that accurately forecast future outcomes.

In performative settings, predictions can actively shape the data. Therefore, we might not only care about accuracy, but also *steering* outcomes towards particular targets. For instance, in education, we might not just want to accurately predict which students will graduate from high school on time, but also maximize the likelihood that they do graduate. We might express this desire to steer objectives by including losses such as $\ell(\hat{y}, y) = 1 - y$ (here, $y = 1$ indicates on-time graduation).[3]

We examine this question regarding the choice of loss and how it relates to the overarching goals of prediction in Chapter 5. However, we include this remark here to encourage the reader to think of the loss ℓ in the performative risk as something that can be quite different from classical objectives in learning theory (e.g. 0-1 or squared loss).

Performative Optimality

The performative risk is a way of measuring how well a model is doing according to a specific objective. This definition naturally motivates that concept of a *performatively optimal* predictor, a model that achieves the minimum possible performative risk:

Definition 2.1.3. A predictor $f_{\theta_{\mathrm{PO}}}$ is performatively optimal with respect to a loss function ℓ and a class of models $\{f_\theta : \theta \in \Theta\}$ if the following relationship holds:

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

That is, $\theta_{\mathrm{PO}} = \arg\min_{\theta \in \Theta} \mathrm{PR}(\theta)$.

The following example illustrates the differences between the traditional notion of optimality in supervised learning and performative optima.

Example 2.1.4 (biased coin flip)**.** Consider the task of predicting the outcome of a biased coin flip where the bias of the coin depends on a feature x and the prediction $f_\theta(x)$.

In particular, define $\mathcal{D}(\theta)$ in the following way. The feature x is a 1-dimensional feature supported on $\{\pm 1\}$ and $Y \mid X \sim \mathrm{Bernoulli}(\frac{1}{2} + \mu X + \epsilon \theta X)$ with $\mu \in (0, \frac{1}{2})$ and $\epsilon < \frac{1}{2} - \mu$. Assume that the class of predictors consists of linear models of the form $f_\theta(x) = \theta x + \frac{1}{2}$ and that the objective is to minimize the squared loss: $\ell(z; \theta) = (y - f_\theta(x))^2$.

[3]Note that this steering loss does not depend explicitly on $\hat{y}$ or θ. The choice of θ influences the performative risk $\mathrm{PR}(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}(\theta)} 1 - y$ only via the performative effects of θ on the distribution over outcomes y.

The parameter ϵ represents the performative aspect of the model. If $\epsilon = 0$, outcomes are independent of the assigned scores and the problem reduces to a standard supervised learning task where the optimal predictive model is the conditional expectation $f_{\theta_{\text{SL}}}(x) = \mathbb{E}[Y \mid X = x] = \frac{1}{2} + \mu x$, with $\theta_{\text{SL}} = \mu$.

In the performative setting with $\epsilon \neq 0$, the optimal model θ_{PO} balances between its predictive accuracy as well as the bias induced by the prediction itself. In particular, a direct calculation demonstrates that

$$\theta_{\text{PO}} = \arg\min_{\theta \in [0,1]} \mathbb{E}_{z \sim \mathcal{D}(\theta)} \left(Y - \theta X - \frac{1}{2} \right)^2 \quad \iff \quad \theta_{\text{PO}} = \frac{\mu}{1 - 2\epsilon}.$$

Hence, the performative optimum and the supervised learning solution are equal if $\epsilon = 0$ and diverge as the performativity strength ϵ increases.

Performative Stability & Retraining

Apart from performative optimality, an alternative, natural property for a model f_θ to satisfy is that, given that we use the predictions of f_θ as a basis for decisions, those predictions are also simultaneously optimal for the distribution that the model induces.

We introduce the notion of *performative stability* to refer to predictive models that satisfy this condition.

Definition 2.1.5 (performative stability and decoupled risk)**.** A model $f_{\theta_{\text{PS}}}$ is *performatively stable* if the following relationship holds:

$$\theta_{\text{PS}} = \arg\min_{\theta} \mathbb{E}_{z \sim \mathcal{D}(\theta_{\text{PS}})} \ell(z; \theta).$$

If we define $\text{DPR}(\theta, \theta') := \mathbb{E}_{z \sim \mathcal{D}(\theta)} \ell(z; \theta')$ as the *decoupled performative risk*; then,

$$\theta_{\text{PS}} = \arg\min_{\theta} \text{DPR}(\theta_{\text{PS}}, \theta).$$

A performatively stable model $f_{\theta_{\text{PS}}}$ minimizes the expected loss on the distribution $\mathcal{D}(\theta_{\text{PS}})$ resulting from deploying $f_{\theta_{\text{PS}}}$ in the first place. Consequently, performatively stable models are the fixed points of repeated retraining.

Repeated retraining, or repeated risk minimization (RRM), is a natural learning dynamic that arises in practice whenever predictions are performative. A learner deploys a model f_θ, and after deployment, realizes that its performance has degraded since the distribution has changed (due to the performative effects of prediction). As a heuristic way to combat this distribution shift, the learner fits a new model to the observed data and redeploys. More formally:

Definition 2.1.6 (RRM)**.** *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 $t \geq 0$:

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

A model that is performatively stable eliminates the need for retraining since any retraining procedure would simply return the same model parameters. This definition and its connection to retraining invites a number of exciting questions.

First of all, unlike performative optimality, stability is a fixed point definition. Therefore, it not at all clear that these stable solutions even exist! There is the possibility that once we begin retraining, we might continue doing so forever without reaching a fixed point. Furthermore, even if there do exist, can we guarantee that they are unique? Lastly, what can we say about their performative risk? Are stable points ever close to performative optima? We will answer all of these over the following chapters.

Before moving on, we observe that performative optimality and performative stability are in general two distinct solution concepts. Performatively optimal models need not be performatively stable and performatively stable models need not be performatively optimal. We illustrate this point in the context of our previous biased coin toss example.

Example 2.1.4 (continued). Consider again our model of a biased coin toss. In order for a predictive model f_θ to be performatively stable, it must satisfy the following relationship:

$$\theta_{\text{PS}} = \arg\min_{\theta \in [0,1]} \mathbb{E}_{Z \sim \mathcal{D}(\theta_{\text{PS}})} \left(Y - \theta X - \frac{1}{2} \right)^2 \quad \iff \quad \theta_{\text{PS}} = \frac{\mu}{1 - \epsilon}.$$

Solving for θ_{PS} directly, we see that there is a unique performatively stable point.

Therefore, performative stability and performative optimality need not identify. In fact, in this example, they identify if and only if $\epsilon = 0$. Note that, in general, if the distribution map $\mathcal{D}(\theta)$ is constant across θ, performative optima must coincide with performatively stable solutions. Furthermore, both coincide with "static" supervised learning solutions as well.

For ease of presentation, we refer to a choice of parameters θ as performatively stable (optimal) if the model parametrized by θ, f_θ is performatively stable (optimal). We will occasionally also refer to performative stability as simply stability.

Remark 2.1.7. Performative stability and optimality can be expressed via the decoupled performative risk as follows:

$$\theta_{\mathrm{PS}} \text{ is performatively stable} \quad \Leftrightarrow \quad \theta_{\mathrm{PS}} = \arg\min_{\theta} \ \mathrm{DPR}(\theta_{\mathrm{PS}}, \theta),$$

$$\theta_{\mathrm{PO}} \text{ is performatively optimal} \quad \Leftrightarrow \quad \theta_{\mathrm{PO}} = \arg\min_{\theta} \ \mathrm{DPR}(\theta, \theta).$$

2.2 Contrasting Optimality and Stability

As discussed previously, performative stability and performative optimality are in general distinct solution concepts. Consequently, they can differ significantly in terms of their relevant performative risks. In this section, we aim to contextualize exactly how different stability and optimality can be, and start to see when we might prefer one vs another.

More concretely, the main goal of this subsection is to set the stage for later algorithmic results. As we will later see, different algorithms lead to different solutions. Repeated retraining of machine learning models leads to *performative stability*. And, depending on the strength of performative effects, these stable solutions may or may not be approximately performatively optimal. By understanding when stability why and optimality differ, we can begin to understand when simple algorithmic solutions, like repeated retraining, achieve good predictive performance, versus when they do not.

We begin by introducing several technical conditions that are relevant for this analysis. One of the core assumptions of the performativity framework is that similar predictive models induce similar distributions. The intuition behind this assumption is quite natural. If two models make very similar predictions, then any subsequent decisions will also be quite similar, and hence induce similar distributions. More formally, we assume that the distribution map $\mathcal{D}(\cdot)$ is a Lipschitz function of the model parameters θ. We refer to this condition as ϵ-sensitivity:

Definition 2.2.1 (ϵ-sensitivity)**.** We say that a distribution map $\mathcal{D}(\cdot)$ is *ϵ-sensitive* if for all $\theta, \theta' \in \Theta$:

$$W_1\big(\mathcal{D}(\theta), \mathcal{D}(\theta')\big) \leq \epsilon \|\theta - \theta'\|_2, \tag{2.2}$$

where W_1 denotes the Wasserstein-1 distance, or earth mover's distance and $\|\cdot\|_2$ denotes the Euclidean norm.

The earth mover's distance is a natural notion of distance between probability distributions that provides access to a rich technical repertoire [85, 86]. Furthermore, we can verify that it is satisfied in various settings.

Example 2.2.2. A simple example where this assumption is satisfied is for a Gaussian family. Given $\theta = (\mu, \sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{2p}$, define $\mathcal{D}(\theta) = \mathcal{N}(\epsilon_1 \mu, \epsilon_2^2 \operatorname{diag}(\sigma_1^2, \ldots, \sigma_p^2))$ where $\epsilon_1, \epsilon_2 \in \mathbb{R}$. Then $\mathcal{D}(\cdot)$ is ϵ-sensitive for $\epsilon = \max\left\{|\epsilon_1|, |\epsilon_2|\right\}$.

Example 2.1.1 (continued). The following, stylized data generating process for the financial trading example from before is also ϵ-sensitive. Assume we predict oil prices as a linear function of some features x that are not affected by prediction. Since outcomes are performative, we might imagine the following model for the distribution map:

$$(x, y) \sim \mathcal{D}(\theta) \iff (x_{\text{base}}, y_{\text{base}}) \sim \mathcal{D}_{\text{base}} \text{ and } (x, y) = \left(x_{\text{base}},\ y_{\text{base}} + \epsilon' \cdot (\theta^\top x_{\text{base}} - y_{\text{base}})\right)$$

That is, there is some base distribution over features and outcomes $\mathcal{D}_{\text{base}}$. In a type of self-fulfilling prophecy, if we make a prediction $\hat{y} = \theta^\top x$, the true prices are nudged closer to our predicted prices by a factor of $\epsilon \cdot (\hat{y} - y_{\text{base}})$, since trading based off the predictions moves the true prices. If features have bounded norm, $\sup_x \|x\|_2 \leq B$, then $\mathcal{D}(\cdot)$ is ϵ sensitive for $\epsilon = \epsilon' B$.

In addition to ϵ-sensitivity, we will make repeated use of the following definitions throughout the remainder of our presentation. To facilitate readability, we let

$$\mathcal{Z} := \cup_{\theta \in \Theta} \operatorname{supp}(\mathcal{D}(\theta)).$$

We say that a loss function $\ell(z; \theta)$ is γ-*strongly convex* in θ if for all $\theta, \theta' \in \Theta$ and $z \in \mathcal{Z}$,

$$\ell(z; \theta) \geq \ell(z; \theta') + \nabla_\theta \ell(z; \theta')^\top (\theta - \theta') + \frac{\gamma}{2} \|\theta - \theta'\|_2^2 . \tag{2.3}$$

If this holds only for $\gamma = 0$, we say that a function is simply convex, or weakly convex. A loss function $\ell(z; \theta)$ is L_z-*Lipschitz* in z if for all $\theta \in \Theta$ and $z, z' \in \mathcal{Z}$,

$$|\ell(z; \theta) - \ell(z'; \theta)| \leq L_z \|z - z'\|_2. \tag{2.4}$$

Likewise, ℓ is L_θ-*Lipschitz* in θ if for all $z \in \mathcal{Z}$ and $\theta, \theta' \in \Theta$,

$$|\ell(z; \theta) - \ell(z; \theta')| \leq L_\theta \|\theta - \theta'\|_2. \tag{2.5}$$

The first result of this section shows that, under appropriate regularity conditions on the loss function, if $\mathcal{D}(\cdot)$ is ϵ-sensitive, all performative optima and stable points lie in a ball of radius at most $O(\epsilon)$.

Recall that the value of ϵ in the definition of ϵ-sensitivity is a measure how strongly predictions can influence the observed data distribution. Small values of ϵ indicate that the data changes very weakly, and smoothly, as a function of our predictions. On the other hand, larger values allow for different models to induce vastly different distributions.

Theorem 2.2.3. *Suppose that the loss $\ell(z;\theta)$ is L_z-Lipschitz in z* (2.4)*, γ-strongly convex in θ* (2.3)*, and that the distribution map $\mathcal{D}(\cdot)$ is ϵ-sensitive* (2.2)*. Then, for every performatively stable point θ_{PS} and every performative optimum θ_{PO}:*

$$\|\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}\|_2 \leq \frac{2L_z\epsilon}{\gamma}.$$

Proof. By definition of performative optimality and performative stability we have that:

$$\mathrm{DPR}(\theta_{\mathrm{PO}}, \theta_{\mathrm{PO}}) \leq \mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PS}}) \leq \mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PO}}).$$

We claim that $\mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PO}}) - \mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PS}}) \geq \frac{\gamma}{2}\|\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}\|_2^2$. By definition of DPR, we can write

$$\mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PO}}) - \mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PS}}) = \mathbb{E}_{Z\sim\mathcal{D}(\theta_{\mathrm{PS}})}\big[\ell(Z;\theta_{\mathrm{PO}}) - \ell(Z;\theta_{\mathrm{PS}})\big].$$

Since $\ell(z;\theta_{\mathrm{PO}}) \geq \ell(z;\theta_{\mathrm{PS}}) + \nabla_\theta \ell(z;\theta_{\mathrm{PS}})^\top(\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}) + \frac{\gamma}{2}\|\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}\|_2^2$ for all z, we have that

$$\begin{aligned}\mathbb{E}_{Z\sim\mathcal{D}(\theta_{\mathrm{PS}})}\big[\ell(Z;\theta_{\mathrm{PO}}) - \ell(Z;\theta_{\mathrm{PS}})\big] &\geq \mathbb{E}_{Z\sim\mathcal{D}(\theta_{\mathrm{PS}})}\big[\nabla_\theta \ell(Z;\theta_{\mathrm{PS}})^\top(\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}})\big] \\ &\quad + \frac{\gamma}{2}\|\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}\|_2^2. \end{aligned} \tag{2.6}$$

Now, by the first order optimality conditions for convex functions,

$$\mathbb{E}_{Z\sim\mathcal{D}(\theta_{\mathrm{PS}})}\big[\nabla_\theta \ell(Z;\theta_{\mathrm{PS}})^\top(\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}})\big] \geq 0,$$

so we get that equation (2.6) implies that:

$$\mathbb{E}_{Z\sim\mathcal{D}(\theta_{\mathrm{PS}})}\big[\ell(Z;\theta_{\mathrm{PO}}) - \ell(Z;\theta_{\mathrm{PS}})\big] \geq \frac{\gamma}{2}\|\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}\|_2^2.$$

Since the population distributions are ϵ-sensitive and the loss is L_z-Lipschitz in z, we have that $\mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PO}}) - \mathrm{DPR}(\theta_{\mathrm{PO}}, \theta_{\mathrm{PO}}) \leq L_z\epsilon\|\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}\|_2$. If $\epsilon < \frac{\gamma\|\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}\|_2}{2L_z}$ then we have that $L_z\epsilon\|\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}\|_2 < \frac{\gamma}{2}\|\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}\|_2^2$ which is a contradiction since it must hold that

$$\mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PO}}) - \mathrm{DPR}(\theta_{\mathrm{PO}}, \theta_{\mathrm{PO}}) \geq \mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PO}}) - \mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PS}}).$$

□

The main message of this theorem is that in cases where the performative effects are small, performatively stable solutions are not all that different from performative optima. This makes intuitive sense. If $\epsilon = 0$, the distribution map is a constant function. In this case, performative prediction reduces to supervised learning since data is drawn from a fixed distribution $\mathcal{D}$ for all models f_θ. As noted in Example 2.1.4, if there is no

performativity, performative optima, performatively stable points, and the supervised learning optima all identify. What this theorem shows is that the extent to which these concepts differ degrades smoothly with the strength of performativity.

Furthermore, under additional regularity conditions on the loss, we can show that performatively stable points are not only close in parameter space to performative optima, they also approximately minimize the performative risk:

Corollary 2.2.4. *Assume that the loss function ℓ is L_z- and L_θ-Lipschitz in z* (2.4) *and θ* (2.5) *respectively, and γ-strongly convex in θ* (2.3)*. If the distribution map is ϵ-sensitive* (2.2)*, then any performative optimal model θ_{PO} and performatively stable solution θ_{PS} satisfy:*

$$\mathrm{PR}(\theta_{\mathrm{PS}}) - \mathrm{PR}(\theta_{\mathrm{PO}}) \leq \frac{2L_z\epsilon(L_\theta + L_z\epsilon)}{\gamma}.$$

Proof. The proof follows from applying Theorem 2.2.3 and the dual formulation of the earth mover's distance (i.e Kantorovich-Rubinstein, Lemma 3.5.1):

$$\begin{aligned}
\mathrm{PR}(\theta_{\mathrm{PS}}) - \mathrm{PR}(\theta_{\mathrm{PO}}) &\leq \left|\mathrm{PR}(\theta_{\mathrm{PS}}) - \mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PO}})\right| + \left|\mathrm{DPR}(\theta_{\mathrm{PS}}, \theta_{\mathrm{PO}}) - \mathrm{PR}(\theta_{\mathrm{PO}})\right| \\
&\leq L_\theta\|\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}\| + L_z\epsilon\|\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}\| \\
&\leq \frac{2L_z\epsilon(L_\theta + L_z\epsilon)}{\gamma}.
\end{aligned}$$

□

Remark 2.2.5. Note that Corollary 2.2.4 and Theorem 2.2.3 hold for *any* pair of performative optima θ_{PO} and performatively stable models θ_{PS}. Neither of these needs to be unique. It could also be the case that performatively stable points do not even exist, in which case the results are vacuously true.

However, this idea that stable points are nearly performatively optimal is only true if performative effects are very weak. That is, ϵ is vanishingly small relative to all other problem parameters. If ϵ is small, but not vanishingly small, Theorem 2.2.3 can be vacuous as we will now see.

Proposition 2.2.6. *For any γ, Δ, and $\epsilon > 0$, there exists a performative prediction problem where the distribution map is ϵ-sensitive and the loss is Lipschitz in both z and θ, as well as γ-strongly convex in θ and smooth in z. Yet, the unique stable point θ_{PS} maximizes the performative risk and*

$$\mathrm{PR}(\theta_{\mathrm{PS}}) - \min_\theta \mathrm{PR}(\theta) \geq \Delta.$$

Proof. Let $z \sim \mathcal{D}(\theta)$ be a point mass at $\epsilon\theta$, and define the loss to be:

$$\ell(z;\theta) = -\beta \cdot \theta^\top z + \frac{\gamma}{2}\|\theta\|_2^2,$$

for some $\beta \geq 0$. This loss is γ-strongly convex and the distribution map is ϵ-sensitive. A short calculation shows that the performative risk simplifies to

$$\mathrm{PR}(\theta) = \left(\frac{\gamma}{2} - \epsilon\beta\right) \cdot \|\theta\|_2^2. \tag{2.7}$$

For $\epsilon \neq \gamma/\beta$, there is a unique performatively stable point at the origin ($\theta = 0$). For β large enough, $\epsilon > \frac{\gamma}{2\beta}$ this point is the unique maximizer of the performative risk. Moreover, for $\epsilon > \frac{\gamma}{2\beta}$, $\min_\theta \mathrm{PR}(\theta) = (\gamma/2 - \epsilon\beta) \cdot \max_{\theta\in\Theta} \|\theta\|_2^2$. Therefore, depending on the radius of Θ, the suboptimality gap of θ_{PS} can be arbitrarily large. □

While the loss is Lipschitz in the counterexample above, the Lipschitz constant scales the diameter of the parameter space Θ. In particular, the Lipschitz constant L_z is equal to $\beta \cdot \max_{\theta\in\Theta} \|\theta\|_2$. Therefore, Theorem 2.2.3 states that stable points and optima are at distance at most $\frac{2L_z\epsilon}{\gamma} = \frac{2\beta\epsilon}{\gamma} \max_{\theta\in\Theta} \|\theta\|_2$. When $\epsilon > \frac{\gamma}{2\beta}$, as assumed in the proof of Proposition 2.2.6, this bound on the distance becomes vacuous: $\|\theta_{\mathrm{PS}} - \theta_{\mathrm{PO}}\|_2 \leq \max_{\theta\in\Theta} \|\theta\|_2$.

Remark 2.2.7. Not that for $\gamma/(2\beta) < \epsilon < \gamma/\beta$, the performative risk is a *concave* function in θ, $\mathrm{PR}(\theta) = -c\|^2\|_2$ for some $c > 0$. Hence, the performative risk can be non-convex even if the loss ℓ is smooth and strongly convex with ϵ smaller than the inverse condition number γ/β.

Conclusion.

Summarizing, in this chapter we have defined the main elements of the performative prediction framework, such as the distribution map $\mathcal{D}(\cdot)$ and the performative risk $\mathrm{PR}(\cdot)$. Furthermore, we've defined the two main solution concepts: performative stability and performative optimality. Performative stability is a fixed point definition whereby models minimize expected risk over the distribution they induce. Whereas, performative optimality is, in general, a distinct concept guaranteeing that a predictor achieves minimal performative risk. If the loss functions are well-conditioned, and the performative effects are vanishingly small, then these solutions achieve similar predictive performance. However, if performative effects are lower bounded by a constant, stability comes with no general guarantees: stable solutions can in fact maximize the performative risk.

2.3 Chapter Notes

The performative prediction framework was introduced in [65] where the authors defined the main concepts and introduced the major theorems relating stability and optimality. Proposition 2.2.6 was presented in [59] as motivation for studying algorithms for finding performatively optimal solutions.

Since its introduction, the ideas behind performative prediction have found several interesting applications. Hardt et al. use the performative prediction lens to study notions of market power in economics. As part of their analysis, they provide an interesting decomposition of the performative risk of a predictive model into terms that represent forecasting and steering, where steering captures the ability of a firm to influence consumer behavior. Furthermore, Malik [52] illustrates how prediction algorithms in the housing market can be performative and induce feedback loops in real estate prices. Mandal et al. [53] extend these ideas to reinforcement learning to formalize the notion that predictions can change the underlying transition to dynamics of a Markov decision process.

Several different works study performativity in a game theoretic context where multiple agents are making predictions simultaneously and jointly influencing the observed data distribution [48, 64, 67]. In biology, [22, 23] establish an exciting connection between performative prediction and feedback design loops that arise when prediction systems are used to study the design of biological sequences (e.g proteins). Lastly, performativity has been used to analyze the kinds of dynamics present in recommendation systems where algorithmic predictions determine the content shown to users and can slowly shape preferences over time [19].

Chapter 3

Understanding Retraining

In the previous chapter, we introduced the concept of the performative risk as a way to measure the value of the prediction rule in settings where predictions are performative and can actively change data distributions. While performative optima are by definition "best-in-class", according to this measure of performance, if predictions are only weakly performative, an alternative solution concept, performative stability, is also near optimal.

This observation has strong implications regarding the validity and impact of common machine learning practices. Often times, people heuristically respond to performative effects of prediction by repeatedly retraining their machine learning models. For example,

- According to their recently open-sourced code, Twitter, a major social media site, trains a model to predict the likelihood that each of their users will interact with a particular tweet [81]. Tweets are ranked according to these predictions and the top few tweets are then displayed to the user.

  These predictions are evidently performative. If the model predicts a user is unlikely to engage with a tweet, they will never observe that piece of content, and hence never click on it. The opposite behavior holds for tweets predicted to be highly relevant.

  Furthermore, these models are "continuously trained" on historical datasets of user-tweet interactions [80]. These datasets were themselves influenced by predictions generated by historical models. This yields a feedback loop whereby the previous model predictions partly determine what the new deployed model will be.

- Google uses supervised learning methods to predict the estimated travel time for people using its maps system [14, 47]. These travel time predictions inform user's decisions regarding which route to take. Consequently, they have the potential to be highly performative [50].

If we predict that a particular route has a relatively low travel time, drivers will be more likely to choose it, hence increasing traffic on this route and altering the true travel time.

As in the Twitter example, these models are periodically retrained on the most recently available traffic data [47]. That is, Google finds the the model that minimizes the empirical risk over the data induced by the previously deployed predictor.

Both of these examples almost exactly match our definition of repeated retraining (Definition 2.1.6):

1. The learner deploys a model θ_t and observes samples drawn $(x, y) \sim \mathcal{D}(\theta_t)$.
2. As a response to the distribution shift, the learner *retrains* and computes:[1]

$$\theta_{t+1} \approx \arg\min_{\theta \in \Theta} \mathbb{E}_{z \sim \mathcal{D}(\theta_t)} \ell(z; \theta).$$

3. The new model θ_{t+1} is deployed and the process starts over again.

The main results of this chapter are algorithmic in nature. Informally speaking, they state that if the performative effects are weak, and well-conditioned, several variants of this natural retraining dynamic are guaranteed to rapidly converge to a stable point.

Taken together, our results provide a new perspective on retraining. If performative effects are a second-order concern (that is predictions only impact the observed data very slightly), retraining is actually a principled way to approach performative prediction problems; this dynamic will quickly converge to an approximately optimal solution. Assuming the relevant loss functions adequately express our social preferences,[2] these results establish how machine learning systems such as the Twitter recommendation algorithm or the Google ETA model are not haphazard heuristics, but theoretically well-founded approaches to social prediction.

[1] We intentionally use $\approx$ to indicate that the exact update rule might vary from problem to problem. However, the key idea is that one tries to minimize expected risk overt the distribution induced by the previous model.

[2] This assumption regarding the choice of loss is crucial. Claims regarding the social benefits of retraining are evidently vacuous if there is a strong mismatch if our mathematical objectives are poor proxies for social value. See discussion in Chapter 5

3.1 Retraining in the Limit of Infinite Data

Exact Retraining

We begin our presentation by considering the simplest and idealized version of repeated risk minimization as seen in Definition 2.1.6. In this setting, the learner deploys a model θ_t and then finds the *exact* population risk risk minimizer,

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

In the real world, the distribution map is not known, hence this algorithm cannot be practically implemented. Nevertheless, it serves as a useful starting point for our analysis of retraining procedures that will guide later results.

As one might imagine, we need to make some assumptions, on both the loss function ℓ and the distribution map $\mathcal{D}(\cdot)$, in order to prove that retraining dynamics are nicely behaved. Otherwise, performatively stable points might not even exist as per the following example:

Example 3.1.1. Consider optimizing the squared loss $\ell(z;\theta) = (y-\theta)^2$, where $\theta \in [0,1]$ and the distribution of the variable y, according to $\mathcal{D}(\theta)$, is a point mass at 0 if $\theta \geq \frac{1}{2}$, and a point mass at 1 if $\theta < \frac{1}{2}$.

Clearly, there is no performatively stable point for this problem, and RRM will simply result in the infinite, alternating sequence $1, 0, 1, 0, \ldots$.

In addition to conditions like ϵ-sensitivity and strong convexity, throughout this chapter we will often assume that loss function is smooth. We say that a loss function ℓ is β_z-smooth in z if the gradient of the loss function with respect to θ is Lipschitz in z.[3] More formally, ℓ is β_z-smooth in z if for all $\theta \in \Theta$:

$$\|\nabla_\theta \ell(z,\theta) - \nabla_\theta \ell(z',\theta)\|_2 \leq \beta_z \|z - z'\|_2. \tag{3.1}$$

Likewise, we say that ℓ is β_θ-smooth in θ if for all z

$$\|\nabla_\theta \ell(z,\theta) - \nabla_\theta \ell(z,\theta')\|_2 \leq \beta_\theta \|\theta - \theta'\|_2. \tag{3.2}$$

Lastly, a loss is β-jointly smooth if it is both β_z-smooth in z and β_θ-smooth in θ for $\beta_z, \beta_\theta \leq \beta$. With these preliminaries out of the way, we are now ready to state our main result regarding the convergence of repeated risk minimization.

[3] Note that even though the Lipschitzness condition is with respect to z, the gradient of the loss ℓ is taken with respect to θ.

Theorem 3.1.2. *Suppose that the loss $\ell(z;\theta)$ is β-smooth in z* (3.1) *and γ-strongly convex in θ* (2.3)*. If the distribution map $\mathcal{D}(\cdot)$ is ϵ-sensitive* (2.2)*, then the following statements are true:*

(a) $\|G(\theta) - G(\theta')\|_2 \leq \epsilon\frac{\beta}{\gamma}\|\theta - \theta'\|_2$, *for all* $\theta, \theta' \in \Theta$.

(b) If $\epsilon < \frac{\gamma}{\beta}$, *the iterates θ_t of RRM converge to a unique performatively stable point θ_{PS} at a linear rate:* $\|\theta_t - \theta_{\text{PS}}\|_2 \leq \delta$ *for all* $t \geq \left(1 - \epsilon\frac{\beta}{\gamma}\right)^{-1} \log\left(\frac{\|\theta_0 - \theta_{\text{PS}}\|_2}{\delta}\right)$.

Proof. Fix $\theta, \theta' \in \Theta$. Let $f(\varphi) = \mathbb{E}_{z\sim\mathcal{D}(\theta)}\ell(z;\varphi)$ and $f'(\varphi) = \mathbb{E}_{z\sim\mathcal{D}(\theta')}\ell(z;\varphi)$. Since f is γ-strongly convex and $G(\theta)$ is the unique minimizer of $f(x)$ we know that,

$$f(G(\theta)) - f(G(\theta')) \geq (G(\theta) - G(\theta'))^\top \nabla f(G(\theta')) + \frac{\gamma}{2}\|G(\theta) - G(\theta')\|_2^2$$
$$f(G(\theta')) - f(G(\theta)) \geq \frac{\gamma}{2}\|G(\theta) - G(\theta')\|_2^2$$

Together, these two inequalities imply that

$$-\gamma\|G(\theta) - G(\theta')\|_2^2 \geq (G(\theta) - G(\theta'))^\top \nabla f(G(\theta')).$$

Next, we observe that $(G(\theta) - G(\theta'))^\top\nabla_\theta\ell(z;G(\theta'))$ is $\|G(\theta) - G(\theta')\|_2\beta$-Lipschitz in z. This follows from applying Cauchy-Schwarz and the fact that the loss is β-smooth in z. Using the dual formulation of the optimal transport distance (Lemma 3.5.1) and ϵ-sensitivity of $\mathcal{D}(\cdot)$,

$$(G(\theta) - G(\theta'))^\top\nabla f(G(\theta')) - (G(\theta) - G(\theta'))^\top\nabla f'(G(\theta')) \geq -\epsilon\beta\|G(\theta) - G(\theta')\|_2\|\theta - \theta'\|_2.$$

Furthermore, using the first-order optimality conditions for convex functions, we have $(G(\theta) - G(\theta'))^\top\nabla f'(G(\theta')) \geq 0$, and hence

$$(G(\theta) - G(\theta'))^\top\nabla f(G(\theta')) \geq -\epsilon\beta\|G(\theta) - G(\theta')\|_2\|\theta - \theta'\|_2.$$

Therefore, we conclude that,

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

Claim (a) then follows by rearranging.

To prove claim (b) we note that $\theta_t = G(\theta_{t-1})$ by the definition of RRM, and $G(\theta_{\text{PS}}) = \theta_{\text{PS}}$ by the definition of stability. Applying the result of part (a) yields

$$\|\theta_t - \theta_{\text{PS}}\|_2 \leq \epsilon\frac{\beta}{\gamma}\|\theta_{t-1} - \theta_{\text{PS}}\|_2 \leq \left(\epsilon\frac{\beta}{\gamma}\right)^t \|\theta_0 - \theta_{\text{PS}}\|_2. \tag{3.3}$$

Setting this expression to be at most δ and solving for t completes the proof of (b). □

Somewhat surprisingly, this convergence result is exactly tight; removing any single assumption required for convergence by Theorem 3.1.2 is enough to construct a counterexample for which RRM diverges:

Proposition 3.1.3. *Suppose that the distribution map $\mathcal{D}(\cdot)$ is ϵ-sensitive with $\epsilon > 0$ (2.2). RRM can fail to converge at all in any of the following cases, for any choice of parameters $\beta, \gamma > 0$:*

(a) The loss is β-jointly smooth ((3.2) & (3.1)) and convex, but not strongly convex in θ (2.3).

(b) The loss is γ-strongly convex in θ, but not jointly smooth.

(c) The loss is β-jointly smooth and γ-strongly convex in θ, but $\epsilon \geq \frac{\gamma}{\beta}$.

Proof. We use a separate counterexample to prove each claim.

Proof of (a) Consider the linear loss defined as $\ell((x, y); \theta) = \beta y \theta$, for $\theta \in [-1, 1]$. Note that this objective is β-smooth in (x, y) and convex in θ, but not strongly convex. Let the distribution of y according to $\mathcal{D}(\theta)$ be a point mass at $\epsilon\theta$, and let the distribution of x be invariant with respect to θ. Clearly, this distribution is ϵ-sensitive.

Here, the decoupled performative risk has the following form $\mathrm{DPR}(\theta, \varphi) = \epsilon\beta\theta\varphi$. The unique performatively stable point is 0. However, if we initialize RRM at any point other than 0, the procedure generates the sequence of iterates $\ldots, 1, -1, 1, -1 \ldots$, thus failing to converge. Furthermore, this behavior holds for all $\epsilon, \beta > 0$.

Proof of (b) Consider a type of regularized hinge loss

$$\ell(z; \theta) = C \max(-1, y\theta) + \frac{\gamma}{2}(\theta - 1)^2,$$

and suppose $\Theta \supseteq [-\frac{1}{2\epsilon}, \frac{1}{2\epsilon}]$.

Let the distribution of Y according to $\mathcal{D}(\theta)$ be a point mass at $\epsilon\theta$, and let the distribution of X be invariant with respect to θ. Clearly, this distribution is ϵ-sensitive. Let $\theta_0 = 2$. Then, by picking C big enough, RRM prioritizes to minimize the first term exactly, and hence we get $\theta_1 = -\frac{1}{2\epsilon}$. In the next step, again due to large C, we get $\theta_2 = 2$. Thus, RRM keeps oscillating between 2 and $-\frac{1}{2\epsilon}$, failing to converge. This argument holds for all $\gamma, \epsilon > 0$.

Proof of (c): Suppose that the loss function is the squared loss, $\ell(z; \theta) = (y - \theta)^2$, where $y, \theta \in \mathbb{R}$. Note that this implies $\beta = \gamma$. Let the distribution of Y according to $\mathcal{D}(\theta)$ be a

point mass at $1 + \epsilon\theta$, and let the distribution of X be invariant with respect to θ. This distribution family satisfies ϵ-sensitivity, because

$$W_1(\mathcal{D}(\theta), \mathcal{D}(\theta')) = \epsilon|\theta - \theta'|_2.$$

By properties of the squared loss, we know

$$\arg\min_{\theta'} \mathrm{DPR}(\theta, \theta') = \mathbb{E}_{Z\sim\mathcal{D}(\theta)}[Y] = 1 + \epsilon\theta.$$

It is thus not hard to see that RRM does not contract if $\epsilon \geq \frac{\gamma}{\beta} = 1$:

$$|G(\theta) - G(\theta')| = |1 + \epsilon\theta - 1 - \epsilon\theta'| = \epsilon|\theta - \theta'|,$$

which exactly matches the bound of Theorem 3.1.2 and proves the first statement of the proposition. The unique performatively stable point of this problem is θ such that $\theta = 1 + \epsilon\theta$, which is $\theta_{\mathrm{PS}} = \frac{1}{1-\epsilon}$ for $\epsilon > 1$.

For $\epsilon = 1$, no performatively stable point exists, thereby proving the second claim of the proposition. If $\epsilon > 1$ on the other hand, and $\theta_0 \neq \theta_{\mathrm{PS}}$, we either have $\theta_t \to \infty$ or $\theta_t \to -\infty$, because

$$\theta_t = 1 + \epsilon\theta_{t-1} = \sum_{k=0}^{t-1} \epsilon^k + \theta_0\epsilon^t = \frac{\epsilon^t - 1}{\epsilon - 1} + \theta_0\epsilon^t,$$

thus concluding the proof. □

This Proposition 3.1.3 leads to a number of interesting conclusion. First, it suggests a fundamental difference between strong and weak convexity in our framing of performative prediction (weak meaning $\gamma = 0$). In supervised learning, using strongly convex losses generally guarantees a faster rate of optimization, yet asymptotically, the solution achieved with either strongly or weakly convex losses is globally optimal. However, in our framework, strong convexity is in fact *necessary* to guarantee convergence of repeated risk minimization, even for arbitrarily smooth losses and an arbitrarily small sensitivity parameter.

Second, this result shows that $\epsilon = \gamma/\beta$ is a sharp threshold which characterizes the convergence of repeated risk minimization. If ϵ is just below the inverse condition number γ/β, then retraining converges. If it's just above, then there are simple counterexample showing it may diverge. However, this does not mean that repeated retraining is guaranteed to diverge on *any* problem for which $\epsilon > \gamma/\beta$. As we will illustrate later on, retraining can still quickly converge to performative stability even for values ϵ that are significantly above this cutoff.[4]

[4]The curious reader may be wondering what happens if ϵ is exactly equal to γ/β. As we will see in later proofs (Proposition 3.1.6), stable points might not exist if ϵ is exactly at the threshold. Hence, we cannot guarantee that RRM will converge to stability.

Gradient Descent & Approximate Retraining

Repeated risk minimization is an elegant, yet highly idealized and unimplementable procedure. Instead of finding the exact risk minimizer on the most recent distribution, a more natural dynamic we might imagine is that the learner simply performs a small update to the model using the new data.

In particular, we can analyze the behavior of repeated gradient descent.

Definition 3.1.4 (RGD)**.** *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 $t \geq 0$:

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

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

Repeated retraining is very much in the spirit of modern optimization algorithms. If loss functions are complicated, we may not be able to compute the optimal solution in closed form. Hence, we sequentially update our solution by following the negative gradient. Furthermore, note that repeated gradient descent only requires the loss ℓ to be differentiable with respect to θ. It does not require taking gradients of the performative risk.

Given that RGD only takes a single gradient step, we might naively expect it to have very different convergence properties to RRM. However, we find that it converges to stability at a comparable (i.e. linear) rate as RRM, and under (nearly) the same conditions. The proofs for the remaining results in this chapter are somewhat technical. We therefore delay them to the end of the chapter.

Proposition 3.1.5. *Assume that the loss ℓ is β-jointly smooth and γ-strongly convex in θ. Furthermore, suppose that the distribution map $\mathcal{D}(\cdot)$ is ϵ-sensitive for $\epsilon < \frac{\gamma}{\beta}$ and that θ_{PS} lies in the interior of the set Θ.*[5] *Then, repeated gradient descent (RGD) with a constant step size $\eta_t = \eta := \frac{\gamma - \epsilon\beta}{2(1+\epsilon^2)\beta^2}$ satisfies the following:*

(a) $\|\theta_{t+1} - \theta_{\mathrm{PS}}\|_2 \leq \left(1 - \frac{\eta(\gamma - \epsilon\beta)}{2}\right) \|\theta_t - \theta_{\mathrm{PS}}\|_2$, *where* $0 < \frac{\eta(\gamma-\epsilon\beta)}{2} < 1$.

(b) The iterates θ_t of RGD converge to the stable point θ_{PS} at a linear rate,

$$\|\theta_{t+1} - \theta_{\mathrm{PS}}\|_2 \leq \delta \text{ for all } t \geq \frac{2}{\eta(\gamma - \epsilon\beta)} \log\left(\frac{\|\theta_0 - \theta_{\mathrm{PS}}\|_2}{\delta}\right).$$

[5] Note for $\epsilon < \gamma/\beta$, Theorem 3.1.2 guarantees that stable point θ_{PS} exists and is unique.

Relative to the analogous result for exact retraining (RRM), Proposition 3.1.5 assumes that the loss is *jointly* smooth in z and θ, whereas we only needed smoothness in z for the previous result. Since RGD is a gradient-based algorithm, we require this additional assumption to ensure that we make sufficient progress with each gradient update. Similar, smoothness type conditions are typical in analyses of gradient methods in convex optimization.

As noted previously, modulo the smoothness assumption, we establish that this convergence analysis of repeated gradient descent is tight:

Proposition 3.1.6. *Suppose that the distribution map $\mathcal{D}(\cdot)$ is ϵ-sensitive* (2.2)*. Repeated gradient descent can fail to converge to a performatively stable point in any of the following cases, for any choice of positive step size sequence $\{\eta_t\}_{t\geq 1}$:*

(a) The loss is β-jointly smooth ((3.2) *&* (3.1)*) and convex in θ* (2.3)*, but not strongly convex, for any $\beta, \epsilon > 0$.*

(b) The loss is β-jointly smooth and γ-strongly convex, but $\epsilon \geq \frac{\gamma}{\beta}$, for any $\gamma, \beta, \epsilon > 0$.

Proof. Let $\Theta = \mathbb{R}$, and let $z \sim \mathcal{D}(\theta)$ be a point mass at $1 + \epsilon\theta$. This distribution map is clearly ϵ-sensitive. Furthermore, define the loss as,

$$\ell(z; \theta) = -\beta z\theta + \frac{\gamma}{2}\theta^2,$$

where $\beta \geq \gamma$ is an arbitrary positive scalar. Note that this objective is convex in θ and β-jointly smooth. Furthermore, it has a unique performatively stable point $\theta_{\text{PS}} = \frac{\beta/\gamma}{1-\epsilon\beta/\gamma}$ whenever $\epsilon \neq \frac{\gamma}{\beta}$; when $\epsilon = \frac{\gamma}{\beta}$, there is no stable point. Repeated gradient descent has the dynamics:

$$\begin{aligned}\theta_{t+1} &= \theta_t - \eta_t \mathbb{E}_{z\sim\mathcal{D}(\theta_t)} \nabla\ell(z; \theta_t) \\ &= \theta_t - \eta_t(\gamma - \epsilon\beta)\theta_t + \eta_t\beta \\ &= (1 - \eta_t(\gamma - \epsilon\beta))\,\theta_t + \eta_t\beta.\end{aligned}$$

If $\gamma = 0$, then the loss $\ell(z; \theta)$ is convex. Furthermore, for any values of $\epsilon, \beta > 0$ and any positive step size sequence $\{\eta_t\}_{k=1}^{\infty}$, it holds that $1 + \eta_t\epsilon\beta > 1$ meaning that RGD diverges.

To prove the second part of the statement, if $\gamma > 0$, then the loss is γ-strongly convex. Furthermore, if $\epsilon > \gamma/\beta$, then for any step size sequence $\{\eta_t\}_{k=1}^{\infty}$, $1 - \eta_t(\gamma - \epsilon\beta) > 1$ and RGD again diverges. When $\epsilon = \frac{\gamma}{\beta}$, there is no stable solution and hence RGD does not converge to stability. □

The main insight from this proposition is that $\epsilon = \gamma/\beta$ is again a sharp threshold for the convergence of RGD, just like it is for RRM. Furthermore, strong convexity is again essential to guarantee any kind of convergence to stability.

Greedy Deploy

Input: step size sequence $\{\eta_k\}_{k=1}^{\infty}$

Deploy initial classifier $\theta_0 \in \Theta$

For each $t = 0, 1, \ldots$

- Observe $z^{(t)} \sim \mathcal{D}(\theta_t)$
- Update model parameters:

$$\theta_{t+1} = \theta_t - \eta_t \nabla \ell(z^{(t)}; \theta_t)$$

- Deploy θ_{t+1}

Lazy Deploy

Input: step size sequence $\{\eta_{t,j}\}_{t,j=1}^{\infty}$

Deploy initial classifier $\theta_0 \in \Theta$

For each $t = 0, 1, \ldots$

- Set $\varphi_{t,1} = \theta_t$
- **For each** $j = 1, \ldots, n(t)$:
  1. Observe $z_j^{(t)} \sim \mathcal{D}(\theta_t)$
  2. Update model parameters:

$$\varphi_{t,j+1} = \varphi_{t,j} - \eta_{t,j} \nabla \ell(z_t^{(t)}; \varphi_{t,j})$$

- Deploy $\theta_{t+1} = \varphi_{k,n(t)}$

Figure 3.1: Stochastic gradient method for performative prediction. Greedy deploy publishes the new model at every step, while lazy deploy performs several gradient updates before releasing the new model.

3.2 Retraining in Finite Samples

Having studied the behavior of various retraining algorithms at the population level, we now move on to analyzing their empirical counterparts which work in finite samples.

More specifically, in this section we study two variants of the stochastic gradient method for optimization in performative settings, which we refer to as *greedy deploy* and *lazy deploy*. At each iteration, both methods use the observed data to perform a stochastic gradient update to the model parameters. However they choose to deploy these updated models at different time intervals.

In the greedy deploy variant, at each iteration we observe a data point drawn from $\mathcal{D}(\theta_t)$, perform a stochastic gradient updates, and immediately redeploy the new model. In practice, data is often plentiful, but model deployments can be quite costly in terms of engineering effort. In such a scenario, it makes sense to aim to minimize the number of model deployment steps by updating the model parameters on multiple data points before initiating another model deployment. This is exactly what the lazy deploy variant

does. The algorithm proceeds in stages. At each stage t, it collects a $n(t) \gg 1$ many samples and uses all of these samples to update the model before redeploying.

Our main theorems for this section are upper bounds on the convergence rate of both of these methods. In particular, we prove that if losses are smooth, strongly convex and the distribution map is ϵ-sensitive with $\epsilon < \gamma/\beta$ then lazy and greedy deploy both asymptotically converge to performative stability.

In terms of their relative performance, the upper bounds on the rates of convergence are nearly identical. However, these are only upper bounds. As such they can only draw an incomplete picture regarding the merits of these methods. We therefore complement our theoretical investigations with empirical simulations. These are presented later on in Section 3.3. Jumping ahead, we find that greedy deploy generally performs better than lazy deploy when the distribution map has a small Lipschitz constant, i.e., the performative effects are small. Conversely, lazy deploy fares better when the distribution map is less Lipschitz.

Greedy Deploy

Before moving onto our analysis, we introduce the following assumption which is customary in the stochastic optimization literature [10, 88].

For the given loss function ℓ, there exist constants σ^2 and L^2 such that for all $\theta, \theta' \in \Theta$:

$$\mathbb{E}_{z\sim\mathcal{D}(\theta)}\left[\|\nabla\ell(z;\theta')\|_2^2\right] \le \sigma^2 + L^2\|\theta' - G(\theta)\|_2^2, \text{ where } G(\theta) := \arg\min_{\theta'} \mathbb{E}_{z\sim\mathcal{D}(\theta)}\ell(z;\theta'). \quad (3.4)$$

We begin by stating a technical lemma which introduces a recursion for the distance between θ_t and θ_{PS}.

Lemma 3.2.1. *Assume that the loss function is β-jointly smooth ((3.2) & (3.1)), γ-strongly convex in θ (2.3) and together with $\mathcal{D}(\cdot)$ satisfies the second moment bound from Equation 3.4. If the distribution map $\mathcal{D}(\cdot)$ is ϵ-sensitive (2.2) with $\epsilon < \gamma/\beta$, then greedy deploy with step size η_t satisfies the following recursion for all $k \ge 1$:*

$$\mathbb{E}\left[\|\theta_{t+1} - \theta_{\mathrm{PS}}\|_2^2\right] \le \left(1 - 2\eta_t(\gamma - \epsilon\beta) + \eta_t^2 L^2\left(1 + \epsilon\frac{\beta}{\gamma}\right)^2\right)\mathbb{E}\left[\|\theta_t - \theta_{\mathrm{PS}}\|_2^2\right] + \eta_t^2\sigma^2.$$

Similar recursions underlie many proofs of SGD, and Lemma 3.2.1 can be seen as their generalization to the performative setting. The key insight achieved by this bound is that implies a strong contraction to the performatively stable point if the performative effects are weak, that is when $\epsilon \ll \gamma/\beta$.

Using this recursion, a simple induction argument suffices to prove that greedy deploy converges to the performatively stable solution. Moreover, it does so at the usual $O(1/k)$ rate.

Theorem 3.2.2. *Assume that the loss function is β-jointly smooth ((3.2) & (3.1)), γ-strongly convex in θ (2.3) and together with $\mathcal{D}(\cdot)$ satisfies the second moment bound from Equation 3.4. If the distribution map $\mathcal{D}(\cdot)$ is ϵ-sensitive (2.2) with $\epsilon < \gamma/\beta$, then for all $k \geq 0$ greedy deploy with step size $\eta_t = \left((\gamma - \epsilon\beta)k + 8L^2/(\gamma - \epsilon\beta)\right)^{-1}$ satisfies*

$$\mathbb{E}\left[\|\theta_{t+1} - \theta_{\mathrm{PS}}\|_2^2\right] \leq \frac{M_{\mathrm{greedy}}}{(\gamma - \epsilon\beta)^2 k + 8L^2},$$

where $M_{\mathrm{greedy}} = \max\left\{2\sigma^2, 8L^2\|\theta_1 - \theta_{\mathrm{PS}}\|_2^2\right\}$.

Comparing this result to the traditional analysis of SGD for smooth, strongly convex objectives (e.g. [69]), we see that the traditional factor of γ is replaced by $\gamma - \epsilon\beta$, which we view as the effective strong convexity parameter of the performative prediction problem. When $\epsilon = 0$, there are no performative effects and the problem of finding the stable solution reduces to that of finding the risk minimizer on a fixed, static distribution. Consequently, it is natural for the two bounds to identify.

Lazy Deploy

Contrary to greedy deploy, lazy deploy collects multiple data points and hence takes multiple stochastic gradient steps between consecutive model deployments. In the Twitter recommendation problem described earlier, this corresponds to only redeploying the prediction model every few weeks, instead of every day.

This modification significantly changes the trajectory of lazy deploy relative to greedy deploy. In particular, the observed samples follow the distribution of the last *deployed* model, which might differ from the current iterate. More precisely, after deploying θ_t, in the lazy deploy variant, we perform $n(t)$ stochastic gradient steps to the model parameters using samples from $\mathcal{D}(\theta_t)$. This yields the sequence of models $\theta_{t,1}, \ldots, \theta_{t,n(t)}$ that are generated offline. At the end of stage t, we deploy the last iterate in this sequence $\theta_{t+1} = \theta_{t,n(t)}$ (see right panel in Figure 3.1).

At a high level, lazy deploy converges to performative stability because it progressively approximates the repeated risk minimization procedure described previously, where

$$\theta_{t+1} = \operatorname*{arg\,min}_{\theta' \in \Theta} \mathbb{E}_{z \sim \mathcal{D}(\theta_t)} \ell(z; \theta').$$

As established in Theorem 3.1.2 converges to a performatively stable classifier at a linear rate when $\epsilon < \gamma/\beta$. Moreover, since the underlying distribution $\mathcal{D}(\theta_t)$remains static

between deployments, a classical analysis of SGD shows that for large values of $n(k)$ these "offline" iterates $\varphi_{k,j}$ converge to the risk minimizer on the distribution corresponding to the previously deployed classifier. In particular, for large $n(k)$, $\theta_{t+1} \approx G(\theta_t)$. By virtue of approximately tracing out the trajectory of RRM, lazy deploy converges to θ_{PS} as well. This intuition is formalized in the following theorem.

Theorem 3.2.3. *Assume that the loss function is β-jointly smooth ((3.2) & (3.1)), γ-strongly convex in θ (2.3) and together with $\mathcal{D}(\cdot)$ satisfies the second moment bound from Equation 3.4. If the distribution map $\mathcal{D}(\cdot)$ is ϵ-sensitive (2.2) with $\epsilon < \gamma/\beta$, for any $\alpha > 0$, running lazy deploy with $n(k) \geq n_0 k^{\alpha}$, $k = 1, 2, \ldots$ many steps between deployments and step size sequence $\eta_{k,j} = (\gamma j + 8L^2/\gamma)^{-1}$, satisfies*

$$\mathbb{E}\left[\|\theta_{t+1} - \theta_{\mathrm{PS}}\|_2^2\right] \leq c^k \cdot \|\theta_1 - \theta_{\mathrm{PS}}\|_2^2 + \left(c^{\Omega(k)} + \frac{2}{k^{\alpha\cdot(1-o(1))}}\right) \cdot M_{\mathrm{lazy}},$$

where $c = \left(\epsilon\frac{\beta}{\gamma}\right)^2 + o(1)$ and $M_{\mathrm{lazy}} = \frac{3(\sigma+\gamma)^2}{\gamma^2(1-c)}$. Here, $o(1)$ is independent of k and vanishes as n_0 grows; n_0 is chosen large enough such that $c < 1$.

Comparing Lazy and Greedy Deploy

As we alluded to previously, the behavior of both algorithms is critically affected by the strength of performative effects ϵ. For $\epsilon \ll \gamma/\beta$, the effective strong convexity parameter $\gamma - \epsilon\beta$ of the performative prediction problem is large. In this setting, the distribution shift induced by deploying a new model is negligible and greedy deploy behaves almost like SGD in classical supervised learning, converging quickly to performative stability.

Conversely, for ϵ close to the convergence threshold, the contraction of greedy deploy to the performatively stable classifier is weak. In this regime, we expect lazy deploy to perform better since the convergence of the offline iterates $\varphi_{k,j}$ to the risk minimizer on the current distribution $G(\theta_k)$ is unaffected by the value of ϵ. Lazy deploy then converges by closely mimicking the behavior of RRM.

In terms of the asymptotics of both algorithms, we identify the following tradeoff between the number of samples and the number of deployments sufficient to converge to performative stability.

Corollary 3.2.4. *Assume that the loss function is β-jointly smooth ((3.2) & (3.1)), γ-strongly convex in θ (2.3) and together with $\mathcal{D}(\cdot)$ satisfies the second moment bound from Equation 3.4. Furthermore, assume that $\mathcal{D}(\cdot)$ is ϵ-sensitive (2.2) with $\epsilon < \gamma/\beta$.*

- *To ensure that greedy deploy returns a solution $\theta^\star$ such that, $\mathbb{E}\left[\|\theta^\star - \theta_{\mathrm{PS}}\|_2^2\right] \leq \delta$, it suffices to collect $O(1 \,/\, \delta)$ samples and to deploy $O(1 \,/\, \delta)$ classifiers.*

- *To achieve the same guarantee using lazy deploy, it suffices to collect* $O(1 \,/\, \delta^{\frac{\alpha+1}{(1-\omega)\cdot\alpha}})$ *samples and to deploy* $O(1 \,/\, \delta^{\frac{1}{\alpha}})$ *classifiers, for any* $\alpha > 0$ *and some* $\omega = 1 - o(1)$ *which tends to 1 as* n_0 *grows.*

We see from the above result that by choosing large enough values of n_0 and α, we can make the sample complexity of the lazy deploy algorithm come arbitrarily close to that of greedy deploy. However, to match the same convergence guarantee, lazy deploy only performs $O(1 \,/\, \delta^{\frac{1}{\alpha}})$ deployments, which is significantly better than the $O(1 \,/\, \delta)$ deployments for greedy deploy.

This reduction in the number of deployments is particularly relevant when considering the settings that performative prediction is meant to address. Whenever we use prediction in social settings, there are important social costs associated with making users adapt to a new model [60]. Furthermore, in industry, there are often significant technical challenges associated with deploying a new classifier. By choosing $n(k) = n_0 k^\alpha$ appropriately, we can reduce the number of deployments necessary for lazy deploy to converge while at the same time improving the sample complexity of the algorithm.

3.3 Applications to Strategic Classification

Having introduced the performative prediction framework, and used it to theoretically analyze various retraining algorithms, in this section:

- Discuss the implications of our theoretical results for strategic classification, a popular framework for learning in social environments.
- Empirically evaluate retraining algorithms on a semi-synthetic simulation environment.

Overview of Strategic Classification

Strategic classification is a two-player game between an institution which deploys a classifier and agents who selectively adapt their features in order to improve their outcomes.

A classic example of this setting is that of a bank which uses a machine learning classifier to predict whether or not a loan applicant is creditworthy. Individual applicants react to the bank's classifier by manipulating their features with the hopes of inducing a

Input: base distribution $\mathcal{D}_{\text{base}}$, classifier f_θ, cost function c, and utility function u

Sampling procedure for $\mathcal{D}(\theta)$:

1. Sample $(x_{\text{base}}, y_{\text{base}}) \sim \mathcal{D}_{\text{base}}$
2. Compute best response $x_{\text{BR}} \leftarrow \arg\max_{x'} u(x', \theta) - c(x', x)$
3. Output sample $(x_{\text{BR}},\ y_{\text{base}})$

Figure 3.2: Distribution map for strategic classification.

favorable classification. This game is said to have a *Stackelberg* structure since agents adapt their features only after the bank has deployed their classifier.[6]

The optimal strategy for the institution in a strategic classification setting is to deploy the solution corresponding to the *Stackelberg equilibrium*, defined as the classifier f_θ which achieves minimal loss over the induced distribution $\mathcal{D}(\theta)$ in which agents have strategically adapted their features in response to f_θ. In fact, we see that this equilibrium notion exactly matches our definition of performative optimality:

$$f_{\theta_{\text{SE}}} \text{ is a Stackelberg equilibrium } \iff \theta_{\text{SE}} \in \arg\min_\theta \text{PR}(\theta).$$

We think of $\mathcal{D}$ as a "baseline" distribution over feature-outcome pairs before any classifier deployment, and $\mathcal{D}(\theta)$ denotes the distribution over features and outcomes obtained by strategically manipulating $\mathcal{D}$. As described in existing work [13, 33, 60], the distribution function $\mathcal{D}(\theta)$ in strategic classification corresponds to the data-generating process outlined in Figure 3.2.

Here, u and c are problem-specific functions which determine the best response for agents in the game. Together with the base distribution $\mathcal{D}$, these define the relevant distribution map $\mathcal{D}(\cdot)$ for the problem of strategic classification.

A strategy that is commonly adapted in practice as a means of coping with the distribution shift that arises in strategic classification is to repeatedly retrain classifiers on the induced distributions. This procedure corresponds to the repeated risk minimization

[6]This is at least the classical framing of strategic classification: the institution moves first and the people being classified respond after. In recent work, Zrnic et al. provide an interesting analysis showing that the exact order of play may be inverted depending on the frequency with which players update their strategies [91].

procedure introduced in Definition 2.1.6, where

$$\theta_{t+1} = \arg\min_{\theta} \mathbb{E}_{z\sim\mathcal{D}(\theta_t)}\ell(z;\theta)$$

Previous theorems regarding the convergence of RRM, RGD, and lazy (or greedy) deploy state that if performative effects are small, and losses are well-conditioned, any one of these variants of retraining will converge to a *performatively stable* point θ_{PS}.

In the context of strategic classification, performative stability corresponds to *Nash equilibria*. Recall that a model θ_{PS} is performatively stable if:

$$\theta_{\text{PS}} = \arg\min_{\theta} \mathbb{E}_{z\sim\mathcal{D}(\theta_{\text{PS}})}\ell(z;\theta).$$

That is, θ_{PS} is the institution's best response to the data distribution that arises from deploying θ_{PS}. Moreover, as seen in Figure 3.2 the data distribution $\mathcal{D}(\theta_{\text{PS}})$ is by definition the people's best response to the classification rule outlined by θ_{PS}. Since both players are playing their best response, we conclude that

$$f_{\theta_{\text{SE}}} \text{ is a Nash equilibrium} \iff \theta_{\text{SE}} \text{ is performatively stable.}$$

Note that by definition, $\text{PR}(\theta_{\text{PO}}) \leq \text{PR}(\theta_{\text{PS}})$. Therefore, Stackelberg equilibria have better performance from the institution's perspective than Nash equilibria (i.e which is the solution one would converge to by repeatedly retraining).

However, in certain cases, Nash equilibria achieves near optimal performative risk. The following corollary is a restatement of Theorem 3.1.2 and Corollary 2.2.4 in the language of strategic classification.

Corollary 3.3.1. *Let the institution's loss $\ell(z;\theta)$ be L_z- and L_θ-Lipschitz in z* (2.4) *and θ* (2.5) *respectively, β-smooth in z* (3.1) *and γ-strongly convex in θ* (2.3)*. If the distribution map is ϵ-sensitive* (2.2)*, with $\epsilon < \frac{\gamma}{\beta}$, then RRM converges at a linear rate to a performatively stable classifier θ_{PS} that is $2L_z\epsilon(L_\theta + L_z\epsilon)\gamma^{-1}$ close in objective value to the Stackelberg equilibrium.*

Taken together, our results describe the first set of sufficient conditions under which repeated retraining overcomes strategic effects.

Simulations

We next examine the convergence of repeated risk minimization and repeated gradient descent in a simulated strategic classification setting. We run experiments on a dynamic credit scoring simulator in which an institution classifies the creditworthiness of loan

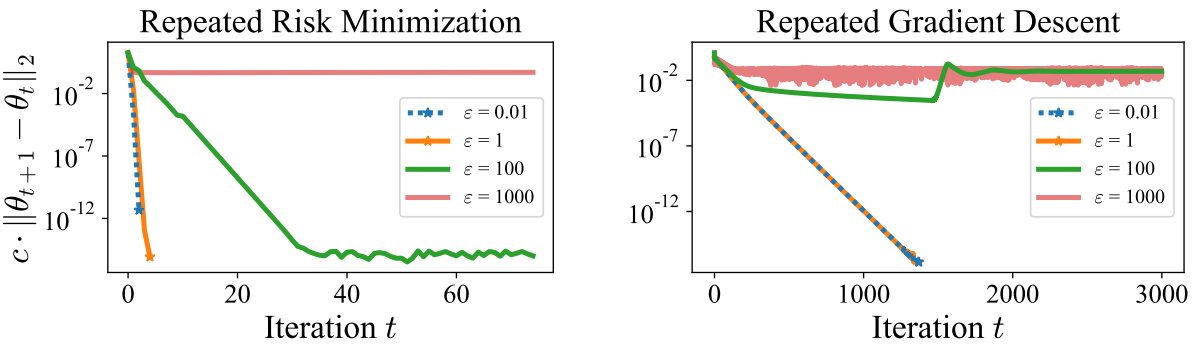


Figure 3.3: Convergence in domain of RRM (left) and RGD (right) for varying ϵ-sensitivity parameters. We add a marker if at the next iteration the distance between iterates is numerically zero. We normalize the distance by $c = \|\theta_{0,S}\|_2^{-1}$.

applicants.[7] As motivated previously, agents react to the institution's classifier by manipulating their features to increase the likelihood that they receive a favorable classification.

To run our simulations, we construct a distribution map $\mathcal{D}(\theta)$, as described in Figure 3.2. For the base distribution $\mathcal{D}_{\text{base}}$, we use a class-balanced subset of a Kaggle credit scoring dataset [40]. Features $x \in \mathbb{R}^{m-1}$ correspond to historical information about an individual, such as their monthly income and number of credit lines. Outcomes $y \in \{0, 1\}$ are binary variables which are equal to 1 if the individual defaulted on a loan and 0 otherwise.

The institution makes predictions using a logistic regression classifier. We assume that individuals have linear utilities $u(\theta, x) = -\langle \theta, x \rangle$ and quadratic costs $c(x', x) = \frac{1}{2\epsilon}\|x' - x\|^2$, where ϵ is a positive constant that regulates the cost incurred by changing features. Linear utilities indicate that agents wish to minimize their assigned probability of default.

We divide the set of features into strategic features $S \subseteq [m-1]$, such as the number of open credit lines, and non-strategic features (e.g., age). Solving the optimization problem described in Figure 3.2, the best response for an individual corresponds to the following update,

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

where $x_S, x'_S, \theta_S \in \mathbb{R}^{|S|}$. As per convention in the literature [13, 33, 60], individual outcomes y are unaffected by strategic manipulation.

[7] Code is available at https://github.com/zykls/performative-prediction, and the simulator has been integrated into the WhyNot software package [58].

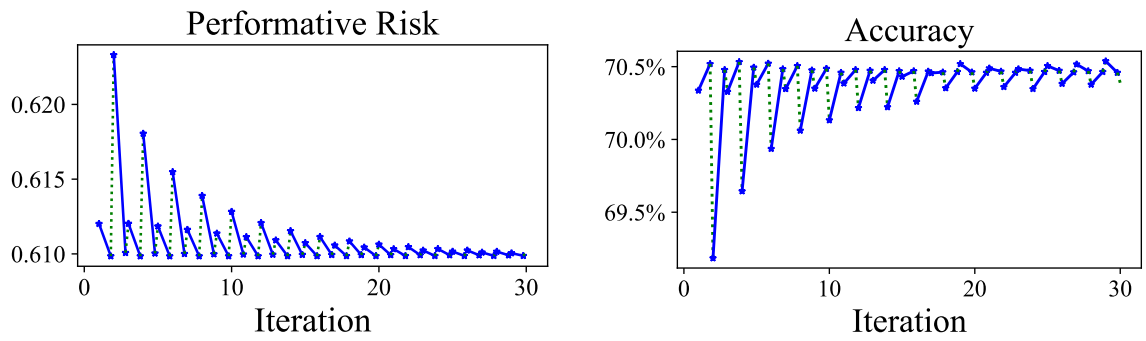


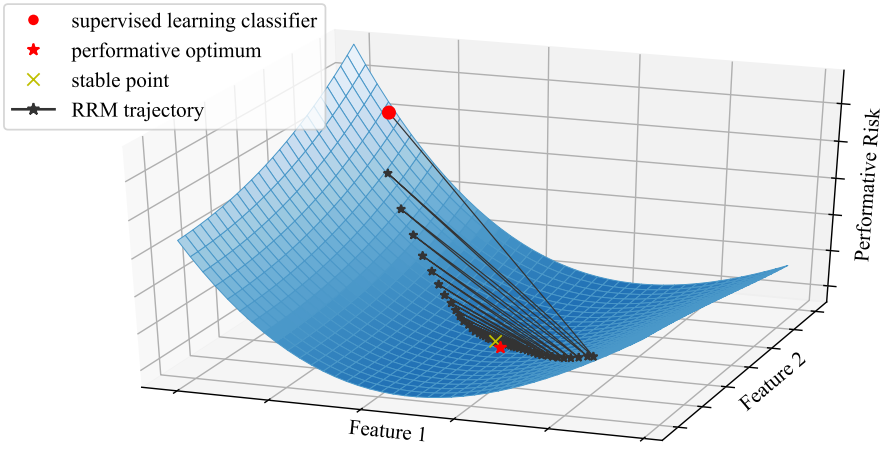
Figure 3.4: Performative risk (left) and accuracy (right) of the classifier θ_t at different stages of RRM for $\epsilon = 80$. Blue lines indicates the optimization phase and green lines indicate the effect of the distribution shift after the classifier deployment.

Intuitively, this data-generating process is ϵ-sensitive since for a given choice of classifiers, f_θ and $f_{\theta'}$, an individual feature vector is shifted to $x_S - \epsilon\theta_S$ and to $x_S - \epsilon\theta'_S$, respectively. The distance between these two shifted points is equal to $\epsilon\|\theta_S - \theta'_S\|_2$. Since the optimal transport distance is bounded by $\epsilon\|\theta - \theta'\|_2$ for every individual point, it is also bounded by this quantity over the entire distribution. In addition, to ϵ-sensitivity, this objective is also jointly smooth. Proofs for of both of these claims are presented in the supplementary material for this chapter. Lastly, we add a regularization term to the logistic loss to ensure that the objective is strongly convex.

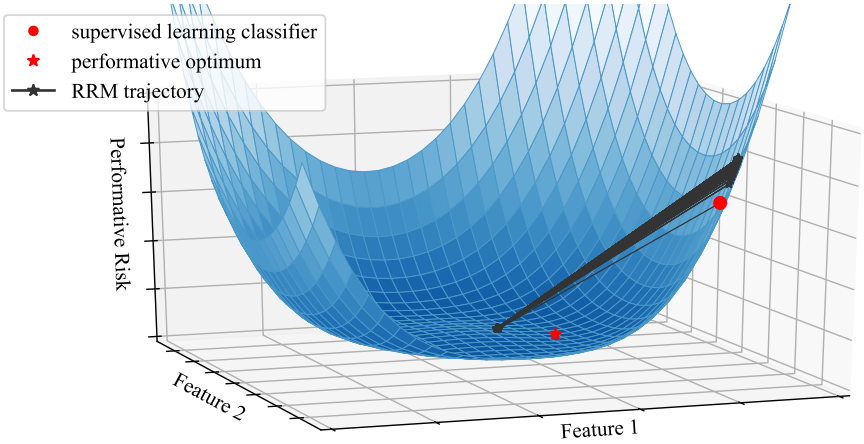
For our experiments with RRM and RGD, instead of sampling from $\mathcal{D}(\theta)$, we treat the points in the original dataset as the true distribution. Hence, we can think of all these procedures as operating at the population level.

Repeated Risk Minimization. The first experiment we consider is the convergence of RRM. From our theoretical analysis, we know that RRM is guaranteed to converge at a linear rate to a performatively stable point if the sensitivity parameter ϵ is smaller than $\frac{\gamma}{\beta}$. In Figure 3.3 (left), we see that RRM does indeed converge in only a few iterations for small values of ϵ while it divergences if ϵ is too large.

The evolution of the performative risk during the RRM optimization is illustrated in Figure 3.4. We evaluate $\mathrm{PR}(\theta)$ at the beginning and at the end of each optimization round and indicate the effect due to distribution shift with a dashed green line. We also verify that the surrogate loss is a good proxy for classification accuracy in the performative setting.



(a) $\epsilon = 25$



(b) $\epsilon = 100$

Figure 3.5: Visualizing the Performative risk surface and trajectory of repeated risk minimization for two different values of sensitivity parameter ϵ. The initial iterate is the risk minimizer on the base dataset (●). We mark the performative optimum (★) and performatively stable point (×).

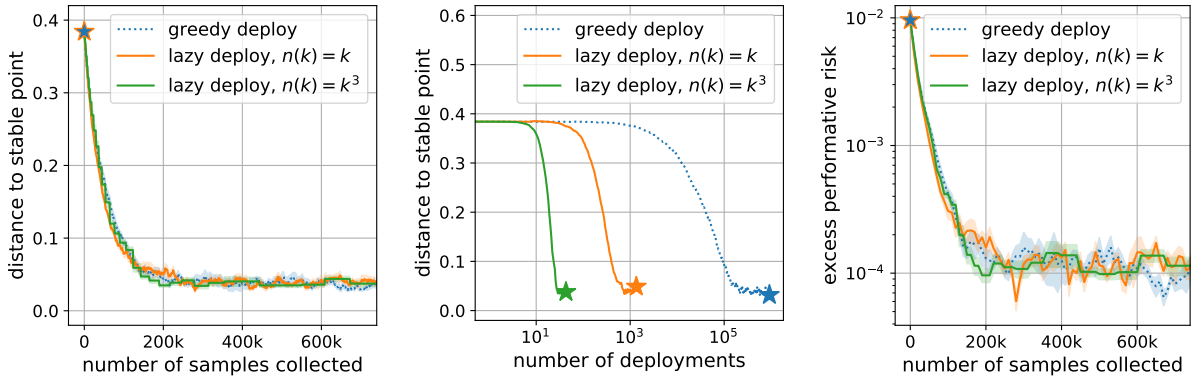


Figure 3.6: Convergence of lazy and greedy deploy to performative stability. Results are for the strategic classification experiments with $\epsilon = 0.001$. (left panel) convergence as a function of the number of samples collected. (center panel) convergence as a function of the number of deployments. (right panel) excess performative risk with respect the the stable classifier θ_{PS} as a function of stochastic gradient updates.

Repeated Gradient Descent. In the case of RGD, we find similar behavior to that of RRM. While the iterates again converge linearly, they naturally do so at a slower rate than in the exact minimization setting, given that each iteration consists only of a single gradient step. Again, we can see in Figure 3.3 that the iterates converge for small values of ϵ and diverge for large values.

Lazy & Greedy Deploy For these experiments, at each time step, the learner observes a single sample from the distribution in which the individual's features have been manipulated in response to the most recently deployed classifier. This is in contrast to the previous experiments for RRM and RGD setup where the learner gets to observe the entire distribution of manipulated features at every step. While we cannot compute the stable point analytically in this setting, we can calculate it empirically by running RRM until convergence.

The inverse condition number for this experiment is quite small $\gamma/\beta \approx 10^{-2}$. We fist pick ϵ within the regime of provable convergence, i.e., $\epsilon = 10^{-3}$, and compare the two methods. As expected, for such a small value of ϵ greedy deploy is the preferred method. Results are depicted in Figure 3.6.

We additionally explore the behavior of these algorithms outside the regime of provable convergence with $\epsilon \gg \gamma/\beta$. We choose step sizes for both algorithms as defined in the theoretical analysis with the exception that we ignore the ϵ-dependence in the step size schedule of greedy deploy and choose the same initial step size as for lazy deploy (Theorem 3.2.2). As illustrated in Figure 3.6 (left), lazy significantly outperforms greedy deploy

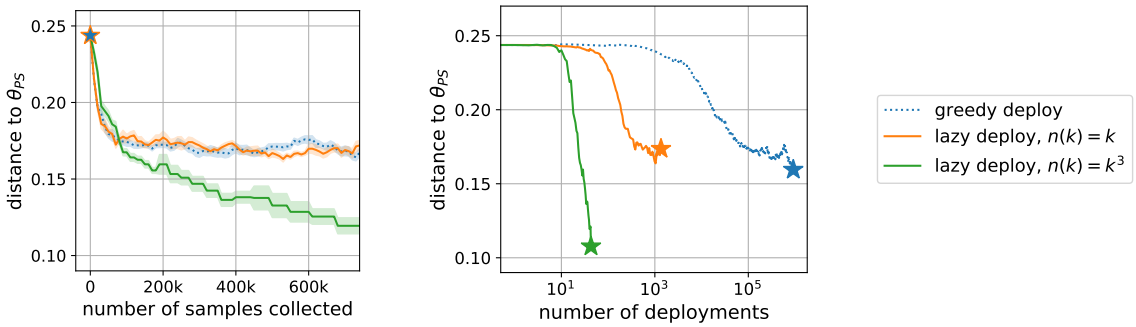


Figure 3.7: Convergence of lazy and greedy deploy to performative stability. Results are for the strategic classification experiments with $\epsilon = 100$. (left panel) convergence as a function of the number of samples. (right panel) convergence as a function of the number of deployments.

in this setting. Moreover, the performance of lazy deploy significantly improves with α. In addition to speeding up convergence, choosing larger sample collection schedules $n(k)$ substantially reduces the number of deployments, as seen in Figure 3.7 (right).

3.4 Chapter Notes

The theoretical results on retraining presented in this chapter first appeared in [65] and [57]. Several of these analyses have since been generalized to hold for other variants of the stochastic gradient method, like proximal point methods or SGD with dual averaging. Drusvyatskiy and Xiao in fact establish a very general and clean analysis of when these variants converge to performative stability. Wood et al. [87] prove complementary results for this stochastic setting.

Throughout our presentation, we have assumed that if the learner deploys θ_t, they immediately see samples drawn from $\mathcal{D}(\theta_t)$. A more realistic assumption is that the distribution maintains a notion of state and that things smoothly vary over time. This idea of including state in performative prediction was first studied by [11] and later extended in [37, 48, 70].

Lastly, while most of the work listed so far studies retraining in linear or convex settings, [61] study the limiting behavior of retraining non-convex neural networks under qualitatively different assumptions.

3.5 Supplementary Material

Technical Lemmas

Lemma 3.5.1 (Kantorovich-Rubinstein)**.** *A distribution map $\mathcal{D}(\cdot)$ is ϵ-sensitive if and only if for all $\theta, \theta' \in \Theta$:*

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

Lemma 3.5.2. *Let $f : \mathbb{R}^n \to \mathbb{R}^d$ be an L-Lipschitz function, and let $X, X' \in \mathbb{R}^n$ be random variables such that $W_1(X, X') \leq C$. Then*

$$\|\mathbb{E}[f(X)] - \mathbb{E}[f(X')]\|_2 \leq LC.$$

Proof.

$$\begin{aligned}\|\mathbb{E}[f(X)] - \mathbb{E}[f(X')]\|_2^2 &= (\mathbb{E}[f(X)] - \mathbb{E}[f(X')])^\top (\mathbb{E}[f(X)] - \mathbb{E}[f(X')]) \\ &= \|\mathbb{E}[f(X)] - \mathbb{E}[f(X')]\|_2 \frac{(\mathbb{E}[f(X)] - \mathbb{E}[f(X')])^\top}{\|\mathbb{E}[f(X)] - \mathbb{E}[f(X')]\|_2} (\mathbb{E}[f(X)] - \mathbb{E}[f(X')]).\end{aligned}$$

Now define the unit vector $v := \frac{\mathbb{E}[f(X)] - \mathbb{E}[f(X')]}{\|\mathbb{E}[f(X)] - \mathbb{E}[f(X')]\|_2}$. By linearity of expectation, we can further write

$$\|\mathbb{E}[f(X)] - \mathbb{E}[f(X')]\|_2^2 = \|\mathbb{E}[f(X)] - \mathbb{E}[f(X')]\|_2 \left(\mathbb{E}[v^\top f(X)] - \mathbb{E}[v^\top f(X')]\right).$$

For any unit vector v and L-Lipschitz function f, $v^\top f$ is a one-dimensional L-Lipschitz function, so we can apply Lemma 3.5.1 to obtain

$$\|\mathbb{E}[f(X)] - \mathbb{E}[f(X')]\|_2^2 \leq \|\mathbb{E}[f(X)] - \mathbb{E}[f(X')]\|_2 LC.$$

Canceling out $\|\mathbb{E}[f(X)] - \mathbb{E}[f(X')]\|_2$ from both sides concludes the proof. □

Lemma 3.5.3 (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

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

Lemma 3.5.4. *Let $s \in (0, 1)$, and fix $\alpha > 0$, then,*

$$\sum_{k=1}^{t} k^{-\alpha} s^{t-k} \leq \frac{s^{t(1-2^{-1/\alpha})}}{1-s} + \frac{2t^{-\alpha}}{1-s}.$$

Proof. Denote by $a_k := k^{-\alpha}$. Let $M_t = \max\{m \in \mathcal{N} : a_m > 2a_t\}$. We decompose the sum depending on M_t as follows:

$$\sum_{k=1}^{t} a_k s^{t-k} = \sum_{k=1}^{M_t} a_k s^{t-k} + \sum_{k=M_t+1}^{t} a_k s^{t-k}.$$

We bound the first term trivially, by applying the fact that $a_k \leq 1$. For the second term, we use the fact that $a_k \leq 2a_t$ for $k > M_t$. We thus get:

$$\sum_{k=1}^{t} a_k s^{t-k} \leq \sum_{k=1}^{M_t} s^{t-k} + 2a_t \sum_{k=M_t+1}^{t} s^{t-k} \leq \frac{s^{t-M_t}}{1-s} + \frac{2a_t}{1-s}.$$

Since $a_k = k^{-\alpha}$, then $M_t \leq \frac{t}{2^{1/\alpha}}$, and so

$$\frac{s^{t-M_t}}{1-s} + \frac{2a_t}{1-s} \leq \frac{s^{t(1-2^{-1/\alpha})}}{1-s} + \frac{2a_t}{1-s}.$$

□

Proof of Lemma 3.2.1

Throughout the proof, we will use $z^{(\theta_{\mathrm{PS}})}$ to denote a sample from $\mathcal{D}(\theta_{\mathrm{PS}})$ which is independent from the whole trajectory of greedy deploy (e.g. $\{\theta_j, z^{(j)}\}_j$, etc.).

Since Θ is closed and convex, we know

$$\|\theta_{k+1} - \theta_{\mathrm{PS}}\|_2^2 = \|\Pi_\Theta(\theta_k - \eta_k \nabla \ell(z^{(k)}; \theta_k)) - \theta_{\mathrm{PS}}\|_2^2 \leq \|\theta_k - \eta_k \nabla \ell(z^{(k)}; \theta_k) - \theta_{\mathrm{PS}}\|_2^2.$$

Squaring the right-hand side and expanding out the square,

$$\begin{aligned}
&\mathbb{E}\big[\|\theta_k - \eta_k \nabla \ell(z^{(k)}; \theta_k) - \theta_{\mathrm{PS}}\|_2^2\big] \\
=\ &\mathbb{E}\big[\|\theta_k - \theta_{\mathrm{PS}}\|_2^2\big] - 2\eta_k \mathbb{E}\big[\nabla \ell(z^{(k)}; \theta_k)^\top (\theta_k - \theta_{\mathrm{PS}})\big] + \eta_k^2 \mathbb{E}\big[\|\nabla \ell(z^{(k)}; \theta_k)\|_2^2\big] \\
:=\ &B_1 - 2\eta_k B_2 + \eta_k^2 B_3.
\end{aligned}$$

We begin by lower bounding B_2. Since θ_{PS} is optimal for the distribution it induces, by Lemma 3.5.3 we have $\mathbb{E}\left[\nabla \ell(z^{(\theta_{\mathrm{PS}})}; \theta_{\mathrm{PS}})^\top (\theta_k - \theta_{\mathrm{PS}})\right] \geq 0$. This allows us to bound B_2 as:

$$\begin{aligned}
B_2 &\geq \mathbb{E}\left[(\nabla \ell(z^{(k)}; \theta_k) - \nabla \ell(z^{(\theta_{\mathrm{PS}})}; \theta_k) + \nabla \ell(z^{(\theta_{\mathrm{PS}})}; \theta_k) - \nabla \ell(z^{(\theta_{\mathrm{PS}})}; \theta_{\mathrm{PS}}))^\top (\theta_k - \theta_{\mathrm{PS}})\right] \\
&= \mathbb{E}\left[(\nabla \ell(z^{(k)}; \theta_k) - \nabla \ell(z^{(\theta_{\mathrm{PS}})}; \theta_k)^\top (\theta_k - \theta_{\mathrm{PS}})\right] \\
&\quad + \mathbb{E}\left[(\nabla \ell(z^{(\theta_{\mathrm{PS}})}; \theta_k) - \nabla \ell(z^{(\theta_{\mathrm{PS}})}; \theta_{\mathrm{PS}}))^\top (\theta_k - \theta_{\mathrm{PS}})\right].
\end{aligned}$$

For the first term, we have that

$$\begin{aligned}
&\mathbb{E}\left[(\nabla\ell(z^{(k)};\theta_k) - \nabla\ell(z^{(\theta_{\text{PS}})};\theta_k)^\top(\theta_k - \theta_{\text{PS}})\right] \\
=\ &\mathbb{E}\Big[\mathbb{E}\left[(\nabla\ell(z^{(k)};\theta_k) - \nabla\ell(z^{(\theta_{\text{PS}})};\theta_k)^\top(\theta_k - \theta_{\text{PS}}) \mid \theta_k\right]\Big] \\
\geq\ &-\epsilon\beta\mathbb{E}\left[\|\theta_k - \theta_{\text{PS}}\|^2\right].
\end{aligned}$$

Having applied the law of iterated expectation, the above inequality follows from the fact that, conditional on θ_k, the function $\nabla\ell(z;\theta_k)^\top(\theta_k - \theta_{\text{PS}})$ is $\beta\|\theta_k - \theta_{\text{PS}}\|-$Lipschitz in z. To verify this claim, we can apply the Cauchy-Schwarz inequality followed by the fact that the gradient is β-jointly smooth. Then, we apply Lemma 3.5.1 and the fact that $\mathcal{D}(\cdot)$ is ϵ-sensitive to get the final bound.

Now, we use strong convexity to bound the second term,

$$\begin{aligned}
&\mathbb{E}\left[(\nabla\ell(z^{(\theta_{\text{PS}})};\theta_k) - \nabla\ell(z^{(\theta_{\text{PS}})};\theta_{\text{PS}}))^\top(\theta_k - \theta_{\text{PS}})\right] \\
=\ &\mathbb{E}\Big[\mathbb{E}\left[(\nabla\ell(z^{(\theta_{\text{PS}})};\theta_k) - \nabla\ell(z^{(\theta_{\text{PS}})};\theta_{\text{PS}}))^\top(\theta_k - \theta_{\text{PS}}) \mid \theta_k\right]\Big] \\
\geq\ &\gamma\mathbb{E}\left[\|\theta_k - \theta_{\text{PS}}\|^2\right].
\end{aligned}$$

Therefore, we get that

$$B_2 \geq (\gamma - \epsilon\beta)\mathbb{E}\left[\|\theta_k - \theta_{\text{PS}}\|^2\right].$$

Now we move on to bounding B_3. Using our assumption on the variance on the gradients yields the following bound, we get

$$\begin{aligned}
\mathbb{E}\left[\|\nabla\ell(z^{(k)};\theta_k)\|_2^2\right] &\leq \sigma^2 + L^2\mathbb{E}\big[\|\theta_k - G(\theta_k)\|^2\big] \\
&= \sigma^2 + L^2\mathbb{E}\big[\|\theta_k - \theta_{\text{PS}} + \theta_{\text{PS}} - G(\theta_k)\|^2\big] \\
&\leq \sigma^2 + L^2\left(\mathbb{E}\big[\left(\|\theta_k - \theta_{\text{PS}}\|_2 + \|\theta_{\text{PS}} - G(\theta_k)\|_2\right)^2\big]\right) \\
&\leq \sigma^2 + L^2\left(1 + \epsilon\frac{\beta}{\gamma}\right)^2\mathbb{E}\big[\|\theta_k - \theta_{\text{PS}}\|^2\big],
\end{aligned}$$

where in the last step we use the fact that G is a contraction mapping (Theorem 3.1.2), which implies $\|\theta_{\text{PS}} - G(\theta_k)\| \leq \epsilon\frac{\beta}{\gamma}\|\theta_k - \theta_{\text{PS}}\|$.

Putting all the steps together completes the proof.

Proof of Proposition 3.1.5

This proof is essentially a consequence of Lemma 3.2.1. By following the steps of Lemma 3.2.1, we get

$$\begin{aligned}\|\theta_{k+1} - \theta_{\mathrm{PS}}\|_2^2 &\leq \|\theta_k - \theta_{\mathrm{PS}}\|_2^2 - 2\eta_k(\mathbb{E}\nabla\ell(z^{(k)};\theta_k))^\top(\theta_k - \theta_{\mathrm{PS}}) + \eta^2\|\mathbb{E}\nabla\ell(z^{(k)};\theta_k)\|_2^2 \\ &:= B_1 - 2\eta B_2 + \eta^2 B_3.\end{aligned}$$

Following the same approach as in Lemma 3.2.1, we get

$$B_2 \geq (\gamma - \epsilon\beta)\|\theta_k - \theta_{\mathrm{PS}}\|_2^2.$$

The bound on B_3 is slightly different, as we no longer make assumptions on the second moment of the gradients; we use $z^{(\theta_{\mathrm{PS}})}$ to denote a sample from $\mathcal{D}(\theta_{\mathrm{PS}})$ and proceed as follows:

$$\begin{aligned}\|\mathbb{E}\nabla\ell(z^{(k)};\theta_k)\|_2^2 &= \|\mathbb{E}\nabla\ell(z^{(k)};\theta_k) - \mathbb{E}\nabla\ell(z^{(\theta_{\mathrm{PS}})};\theta_{\mathrm{PS}})\|_2^2 \\ &\leq \|\mathbb{E}\nabla\ell(z^{(k)};\theta_k) - \mathbb{E}\nabla\ell(z^{(k)};\theta_{\mathrm{PS}}) + \mathbb{E}\nabla\ell(z^{(k)};\theta_{\mathrm{PS}}) - \mathbb{E}\nabla\ell(z^{(\theta_{\mathrm{PS}})};\theta_{\mathrm{PS}})\|_2^2 \\ &\leq 2\|\mathbb{E}\nabla\ell(z^{(k)};\theta_k) - \mathbb{E}\nabla\ell(z^{(k)};\theta_{\mathrm{PS}})\|_2^2 \\ &\quad + 2\|\mathbb{E}\nabla\ell(z^{(k)};\theta_{\mathrm{PS}}) - \mathbb{E}\nabla\ell(z^{(\theta_{\mathrm{PS}})};\theta_{\mathrm{PS}})\|_2^2 \\ &\leq 2\beta^2\|\theta_k - \theta_{\mathrm{PS}}\|_2^2 + 2\beta^2\epsilon^2\|\theta_k - \theta_{\mathrm{PS}}\|_2^2 \\ &\leq 2\beta^2\left(1 + \epsilon^2\right)\|\theta_k - \theta_{\mathrm{PS}}\|_2^2,\end{aligned}$$

where in the third inequality we apply the fact that the loss if β-jointly smooth, together with Lemma 3.5.2. Putting everything together, this implies

$$\|\theta_{k+1} - \theta_{\mathrm{PS}}\|_2^2 \leq (1 - 2\eta(\gamma - \epsilon\beta) + 2\eta^2\beta^2(1 + \epsilon^2))\|\theta_k - \theta_{\mathrm{PS}}\|_2^2.$$

Using the fact that $\sqrt{1-x} \leq 1 - \frac{x}{2}$ for $x \in [0, 1]$, we get

$$\|\theta_{k+1} - \theta_{\mathrm{PS}}\|_2 \leq (1 - \eta(\gamma - \epsilon\beta) + \eta^2\beta^2(1 + \epsilon^2))\|\theta_k - \theta_{\mathrm{PS}}\|_2.$$

By setting $\eta = \frac{\gamma - \epsilon\beta}{2(1+\epsilon^2)\beta^2}$, we can conclude

$$\|\theta_{k+1} - \theta_{\mathrm{PS}}\|_2 \leq \left(1 - \frac{(\gamma - \epsilon\beta)^2}{4(1 + \epsilon^2)\beta^2}\right)\|\theta_k - \theta_{\mathrm{PS}}\|_2.$$

Note that $\frac{(\gamma-\epsilon\beta)^2}{4(1+\epsilon^2)\beta^2} < 1$ because $(\gamma - \epsilon\beta)^2 \leq \gamma^2 + \epsilon^2\beta^2 \leq (1 + \epsilon^2)\beta^2$.

We can unroll the above recursion to get

$$\begin{aligned}\|\theta_{k+1} - \theta_{\text{PS}}\|_2 &\leq \left(1 - \frac{(\gamma - \epsilon\beta)^2}{4(1+\epsilon^2)\beta^2}\right)^k \|\theta_1 - \theta_{\text{PS}}\|_2 \\ &\leq \exp\left(-\frac{k(\gamma - \epsilon\beta)^2}{4(1+\epsilon^2)\beta^2}\right) \|\theta_1 - \theta_{\text{PS}}\|_2.\end{aligned}$$

Setting the right-hand side to δ and expressing k completes the proof.

Proof of Greedy Deploy: Theorem 3.2.2

From Lemma 3.2.1, we have that the following recursion holds:

$$\mathbb{E}\left[\|\theta_{k+1} - \theta_{\text{PS}}\|_2^2\right] \leq \left(1 - 2\eta_k(\gamma - \epsilon\beta) + \eta_k^2 L^2 \left(1 + \epsilon\frac{\beta}{\gamma}\right)^2\right) \mathbb{E}\left[\|\theta_k - \theta_{\text{PS}}\|_2^2\right] + \eta_k^2\sigma^2.$$

Using the fact that $\epsilon < \frac{\gamma}{\beta}$, we get that,

$$\mathbb{E}\left[\|\theta_{k+1} - \theta_{\text{PS}}\|_2^2\right] \leq \left(1 - 2\eta_k(\gamma - \epsilon\beta) + 4\eta_k^2 L^2\right) \mathbb{E}\left[\|\theta_k - \theta_{\text{PS}}\|_2^2\right] + \eta_k^2\sigma^2.$$

We proceed by using induction. As in the theorem statement, we let $\eta_k = \frac{1}{(\gamma-\epsilon\beta)(k+k_0)}$, where we denote $k_0 = \frac{8L^2}{(\gamma-\epsilon\beta)^2}$. The base case, $k = 0$, is trivially true by construction of the bound and choice of k_0. Now, we adopt the inductive hypothesis that

$$\mathbb{E}\left[\|\theta_{k+1} - \theta_{\text{PS}}\|_2^2\right] \leq \frac{\max\left\{2\sigma^2, 8L^2\|\theta_1 - \theta_{\text{PS}}\|_2^2\right\}}{(\gamma - \epsilon\beta)^2(k + k_0)}.$$

Then, by Lemma 3.2.1, it is true that

$$\begin{aligned}\mathbb{E}\left[\|\theta_{k+2} - \theta_{\text{PS}}\|_2^2\right] &\leq \left(1 - 2\eta_k(\gamma - \epsilon\beta) + 4\eta_k^2 L^2\right) \mathbb{E}\left[\|\theta_{k+1} - \theta_{\text{PS}}\|_2^2\right] + \eta_k^2\sigma^2 \\ &\leq \frac{1}{(\gamma - \epsilon\beta)^2}\left(\frac{k + k_0 - 2 + \frac{4L^2}{(\gamma-\epsilon\beta)^2 k_0}}{(k + k_0)^2} \max\left\{2\sigma^2, 8L^2\|\theta_1 - \theta_{\text{PS}}\|_2^2\right\} + \frac{\sigma^2}{(k + k_0)^2}\right) \\ &\leq \frac{1}{(\gamma - \epsilon\beta)^2}\left(\frac{k + k_0 - 1.5}{(k + k_0)^2} \max\left\{2\sigma^2, 8L^2\|\theta_1 - \theta_{\text{PS}}\|_2^2\right\} + \frac{\sigma^2}{(k + k_0)^2}\right) \\ &\leq \frac{1}{(\gamma - \epsilon\beta)^2}\left(\frac{k + k_0 - 1}{(k + k_0)^2} \max\left\{2\sigma^2, 8L^2\|\theta_1 - \theta_{\text{PS}}\|_2^2\right\} - \frac{0.5 \cdot 2\sigma^2 - \sigma^2}{(k + k_0)^2}\right) \\ &= \frac{1}{(\gamma - \epsilon\beta)^2} \cdot \frac{k + k_0 - 1}{(k + k_0)^2} \max\left\{2\sigma^2, 8L^2\|\theta_1 - \theta_{\text{PS}}\|_2^2\right\} \\ &\leq \frac{1}{(\gamma - \epsilon\beta)^2} \cdot \frac{1}{k + 1 + k_0} \max\left\{2\sigma^2, 8L^2\|\theta_1 - \theta_{\text{PS}}\|_2^2\right\},\end{aligned}$$

where the last step follows because $(k+k_0)^2 > (k+k_0)^2 - 1 = (k+k_0+1)(k+k_0-1)$. Therefore, we have shown $\mathbb{E}\left[\|\theta_{k+2} - \theta_{\text{PS}}\|_2^2\right] \leq \frac{M_{\text{greedy}}}{(\gamma-\epsilon\beta)^2(k+1+k_0)}$, which completes the proof by induction.

Proof of Lazy Deploy: Theorem 3.2.3

To prove Theorem 3.2.3, we use the following classical result about convergence of SGD on a static distribution (see, e.g., [69]). The step size is chosen such that it matches the step size of Theorem 3.2.2 when $\epsilon = 0$. We include the proof for completeness.

Lemma 3.5.5. *If the loss is γ-strongly convex in θ* (2.3) *and satisfied* (3.4) *then Lazy deploy satisfies the following:*

$$\mathbb{E}\left[\|\varphi_{k,j+1} - G(\theta_k)\|_2^2\right] \leq \left(1 - 2\eta_{k,j}\gamma + \eta_{k,j}^2 L^2\right)\mathbb{E}\left[\|\varphi_{k,j} - G(\theta_k)\|_2^2\right] + \eta_{k,j}^2\sigma^2.$$

If, additionally, $\eta_{k,j} = \frac{1}{\gamma j + 8L^2/\gamma}$, then for all $k \geq 1$, $j \geq 0$, the following is true

$$\mathbb{E}\left[\|\varphi_{k,j+1} - G(\theta_k)\|_2^2\right] \leq \frac{M_{\text{lazy}}}{\gamma^2 j + L^2},$$

where $M_{\text{lazy}} := \max\left\{1.2\sigma^2, 8L^2\mathbb{E}[\|\theta_k - G(\theta_k)\|_2^2]\right\}$.

Proof. First we prove the recursion. Since Θ is closed and convex, we know

$$\begin{aligned}
&\mathbb{E}\left[\|\varphi_{k,j+1} - G(\theta_k)\|_2^2\right] \\
=\ &\mathbb{E}\left[\left\|\Pi_\Theta\left(\varphi_{k,j} - \eta_{k,j}\nabla\ell(z_j^{(k)};\varphi_{k,j})\right) - G(\theta_k)\right\|_2^2\right] \\
\leq\ &\mathbb{E}\Big[\left\|\varphi_{k,j} - \eta_{k,j}\nabla\ell(z_j^{(k)};\varphi_{k,j}) - G(\theta_k)\right\|_2^2\Big] \\
=\ &\mathbb{E}\Big[\left\|\varphi_{k,j} - G(\theta_k)\right\|_2^2\Big] - 2\eta_{k,j}\mathbb{E}\Big[\nabla\ell(z_j^{(k)};\varphi_{k,j})^\top(\varphi_{k,j} - G(\theta_k))\Big] + \eta_{k,j}^2\mathbb{E}\Big[\|\nabla\ell(z_j^{(k)};\varphi_{k,j})\|^2\Big].
\end{aligned}$$

Next, we examine the cross-term. By the first-order optimality conditions for convex functions (Lemma 3.5.3), we know that $\mathbb{E}\left[\nabla\ell(z_j^{(k)}; G(\theta_k))^\top(\varphi_{k,j} - G(\theta_k))\right] \geq 0$. Using this lemma along with strong convexity, we can lower bound this term as follows,

$$\begin{aligned}
\mathbb{E}\Big[\nabla\ell(z_j^{(k)};\varphi_{k,j})^\top(\varphi_{k,j} - G(\theta_k))\Big] &\geq \mathbb{E}\Big[(\nabla\ell(z_j^{(k)};\varphi_{k,j}) - \nabla\ell(z_j^{(k)}; G(\theta_k))^\top(\varphi_{k,j} - G(\theta_k))\Big] \\
&\geq \gamma\mathbb{E}\Big[\|\varphi_{k,j} - G(\theta_k)\|^2\Big].
\end{aligned}$$

For the final term, we use our assumption on the second moment of the gradients,

$$\mathbb{E}\left[\|\nabla\ell(z_j^{(k)};\varphi_{k,j})\|_2^2\right] \leq \sigma^2 + L^2\mathbb{E}\left[\|\varphi_{k,j} - G(\theta_k)\|_2^2\right].$$

Putting everything together, we get the desired recursion,

$$\mathbb{E}\left[\|\varphi_{k,j+1} - G(\theta_k)\|_2^2\right] \leq (1 - 2\eta_{k,j}\gamma + \eta_{k,j}^2 L^2)\mathbb{E}\left[\|\varphi_{k,j} - G(\theta_k)\|_2^2\right] + \eta_{k,j}^2\sigma^2.$$

Now we turn to proving the second part of the lemma. We prove the result using induction. As in the theorem statement, we let $\eta_{k,j} = \frac{1}{\gamma(j+k_0)}$, where we denote $k_0 = \frac{8L^2}{\gamma^2}$. The base case, $j = 0$, is trivially true by construction of the bound and choice of k_0. Now, we adopt the inductive hypothesis that

$$\mathbb{E}\left[\|\varphi_{k,j+1} - G(\theta_k)\|_2^2\right] \leq \frac{\max\left\{1.2\sigma^2, 8L^2\mathbb{E}\left[\|\theta_k - G(\theta_k)\|_2^2\right]\right\}}{\gamma^2(j + k_0)}.$$

Then, by part (a) of this lemma, it is true that

$$\begin{aligned}
\mathbb{E}\left[\|\varphi_{k,j+2} - G(\theta_k)\|_2^2\right] &\leq \left(1 - 2\eta_{k,j}\gamma + \eta_{k,j}^2 L^2\right)\mathbb{E}\left[\|\varphi_{k,j+1} - G(\theta_k)\|_2^2\right] + \eta_{k,j}^2\sigma^2 \\
&\leq \frac{1}{\gamma^2}\left(\frac{j + k_0 - 2 + \frac{L^2}{\gamma^2 k_0}}{(j + k_0)^2}\max\left\{1.2\sigma^2, 8L^2\mathbb{E}\left[\|\theta_k - G(\theta_k)\|_2^2\right]\right\} + \frac{\sigma^2}{(j + k_0)^2}\right) \\
&\leq \frac{1}{\gamma^2}\left(\frac{j + k_0 - 15/8}{(j + k_0)^2}\max\left\{1.2\sigma^2, 8L^2\mathbb{E}\left[\|\theta_k - G(\theta_k)\|_2^2\right]\right\} + \frac{\sigma^2}{(j + k_0)^2}\right) \\
&\leq \frac{1}{\gamma^2}\left(\frac{j + k_0 - 1}{(j + k_0)^2}\max\left\{1.2\sigma^2, 8L^2\mathbb{E}\left[\|\theta_k - G(\theta_k)\|_2^2\right]\right\} - \frac{7/8 \cdot 1.2\sigma^2 + \sigma^2}{(j + k_0)^2}\right) \\
&= \frac{1}{\gamma^2}\cdot\frac{j + k_0 - 1}{(j + k_0)^2}\max\left\{1.2\sigma^2, 8L^2\mathbb{E}\left[\|\theta_k - G(\theta_k)\|_2^2\right]\right\} \\
&\leq \frac{1}{\gamma^2}\cdot\frac{1}{j + 1 + k_0}\max\left\{1.2\sigma^2, 8L^2\mathbb{E}\left[\|\theta_k - G(\theta_k)\|_2^2\right]\right\},
\end{aligned}$$

where the last step follows because $(j + k_0)^2 > (j + k_0)^2 - 1 = (j + k_0 + 1)(j + k_0 - 1)$. Therefore, we have shown $\mathbb{E}\left[\|\varphi_{k,j+2} - G(\theta_k)\|_2^2\right] \leq \frac{M_{\text{lazy}}}{\gamma^2(j+1+k_0)}$, which completes the proof by induction. □

Now we finally prove Theorem 3.2.3. First we state two deterministic identities used in the proof, which follow from Theorem 3.1.2.

$$\|G(\theta) - \theta_{\text{PS}}\|_2 \leq \epsilon\frac{\beta}{\gamma}\|\theta - \theta_{\text{PS}}\|_2, \tag{3.5}$$

$$\|\theta - G(\theta)\|_2 \leq \|\theta - \theta_{\text{PS}}\|_2 + \|\theta_{\text{PS}} - G(\theta)\|_2 \leq \left(1 + \epsilon\frac{\gamma}{\beta}\right)\|\theta - \theta_{\text{PS}}\|_2. \tag{3.6}$$

Note that identity (3.6) implies $\|\theta - G(\theta)\|_2 < 2\|\theta - \theta_{\text{PS}}\|_2$ if $\epsilon < \frac{\gamma}{\beta}$.

By triangle inequality, we have

$$\begin{aligned}
&\mathbb{E}\left[\|\theta_{k+1} - \theta_{\mathrm{PS}}\|_2^2\right] \\
&= \mathbb{E}\left[\|\theta_{k+1} - G(\theta_k) + G(\theta_k) - \theta_{\mathrm{PS}}\|_2^2\right] \\
&\leq \mathbb{E}\left[\|\theta_{k+1} - G(\theta_k)\|_2^2\right] + 2\mathbb{E}\left[\|\theta_{k+1} - G(\theta_k)\|_2 \|G(\theta_k) - \theta_{\mathrm{PS}}\|_2\right] + \mathbb{E}\left[\|G(\theta_k) - \theta_{\mathrm{PS}}\|_2^2\right].
\end{aligned} \tag{3.7}$$

Denoting $k_0 = \frac{8L^2}{\gamma^2}$, Lemma 3.5.5 bounds the first term by

$$\begin{aligned}
\mathbb{E}\left[\|\theta_{k+1} - G(\theta_k)\|_2^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\|\theta_{k+1} - G(\theta_k)\|_2^2 \mid \theta_k\right]\right] \\
&\leq \frac{1.2\sigma^2 + 8L^2\mathbb{E}\left[\|\theta_k - G(\theta_k)\|_2^2\right]}{\gamma^2(n(k) + k_0)} \\
&\leq \frac{1.2\sigma^2 + 32L^2\mathbb{E}\left[\|\theta_k - \theta_{\mathrm{PS}}\|_2^2\right]}{\gamma^2(n(k) + k_0)},
\end{aligned}$$

where in the last step we apply identity (3.6). Note also that by Jensen's inequality, we know

$$\mathbb{E}\left[\|\theta_{k+1} - G(\theta_k)\|_2\right] \leq \frac{1.1\sigma + 6L\mathbb{E}\left[\|\theta_k - G(\theta_k)\|_2\right]}{\gamma\sqrt{n(k) + k_0}}.$$

We can use this inequality, together with identities (3.5) and (3.6), to bound the cross-term in equation (3.7) as follows:

$$\begin{aligned}
&2\mathbb{E}\left[\|\theta_{k+1} - G(\theta_k)\|_2 \|G(\theta_k) - \theta_{\mathrm{PS}}\|_2\right] \\
&\leq 2\epsilon\frac{\beta}{\gamma}\mathbb{E}\left[\|\theta_{k+1} - G(\theta_k)\|_2 \|\theta_k - \theta_{\mathrm{PS}}\|_2\right] \\
&\leq \frac{2\epsilon\frac{\beta}{\gamma}}{\sqrt{n(k) + k_0}}\mathbb{E}\left[\left(\frac{6L}{\gamma}\|\theta_k - G(\theta_k)\|_2 + \frac{1.1\sigma}{\gamma}\right)\|\theta_k - \theta_{\mathrm{PS}}\|_2\right] \\
&\leq \frac{2\epsilon\frac{\beta}{\gamma}}{\sqrt{n(k) + k_0}}\mathbb{E}\left[\left(\frac{6L}{\gamma}\left(1 + \epsilon\frac{\beta}{\gamma}\right)\|\theta_k - \theta_{\mathrm{PS}}\|_2 + \frac{1.1\sigma}{\gamma}\right)\|\theta_k - \theta_{\mathrm{PS}}\|_2\right] \\
&\leq \frac{24\epsilon\beta L}{\gamma^2\sqrt{n(k) + k_0}}\mathbb{E}\left[\|\theta_k - \theta_{\mathrm{PS}}\|_2^2\right] + \frac{2.2\sigma\epsilon\beta}{\gamma^2\sqrt{n(k) + k_0}}\mathbb{E}\left[\|\theta_k - \theta_{\mathrm{PS}}\|_2\right].
\end{aligned}$$

We bound the latter term by applying the AM-GM inequality; in particular, for all $\alpha_0 \in (0, 1)$, it holds that

$$\frac{2.2\sigma\epsilon\beta}{\gamma^2\sqrt{n(k) + k_0}}\mathbb{E}\left[\|\theta_k - \theta_{\mathrm{PS}}\|_2\right] \leq \frac{1.1\sigma\epsilon\beta}{\gamma^2}\left(\frac{1}{(n(k) + k_0)^{\alpha_0}} + \frac{\mathbb{E}\left[\|\theta_k - \theta_{\mathrm{PS}}\|_2^2\right]}{(n(k) + k_0)^{1-\alpha_0}}\right).$$

Thus, the final bound on the cross-term in equation (3.7) is

$$2\mathbb{E}\left[\|\theta_{k+1} - G(\theta_k)\|_2 \|G(\theta_k) - \theta_{\mathrm{PS}}\|_2\right] \leq \left(\frac{24\epsilon\beta L}{\gamma^2\sqrt{n(k)+k_0}} + \frac{1.1\sigma\epsilon\beta}{\gamma^2(n(k)+k_0)^{1-\alpha_0}}\right)\mathbb{E}\left[\|\theta_k - \theta_{\mathrm{PS}}\|_2^2\right] + \frac{1.1\sigma\epsilon\beta}{\gamma^2(n(k)+k_0)^{\alpha_0}}.$$

The final term in equation (3.7) can be bounded by identity (3.5):

$$\mathbb{E}\left[\|G(\theta_k) - \theta_{\mathrm{PS}}\|_2^2\right] \leq \left(\epsilon\frac{\beta}{\gamma}\right)^2 \mathbb{E}\left[\|\theta_k - \theta_{\mathrm{PS}}\|_2^2\right].$$

Putting all the steps together, we have derived the following recursion, true for all $\alpha_0 \in (0,1)$:

$$\begin{aligned}\mathbb{E}\left[\|\theta_{k+1} - \theta_{\mathrm{PS}}\|_2^2\right] &\leq \left(\frac{32L^2}{\gamma^2(n(k)+k_0)} + \frac{24\epsilon\beta L}{\gamma^2\sqrt{n(k)+k_0}} + \frac{1.1\sigma\epsilon\beta}{\gamma^2(n(k)+k_0)^{1-\alpha_0}} + \left(\epsilon\frac{\beta}{\gamma}\right)^2\right)\mathbb{E}\left[\|\theta_k - \theta_{\mathrm{PS}}\|_2^2\right] \\ &\quad + \frac{1.2\sigma^2}{\gamma^2(n(k)+k_0)} + \frac{1.1\sigma\epsilon\beta}{\gamma^2(n(k)+k_0)^{\alpha_0}} \\ &\leq c\mathbb{E}\left[\|\theta_k - \theta_{\mathrm{PS}}\|_2^2\right] + \frac{1.2\sigma^2}{\gamma^2(n(k)+k_0)} + \frac{1.1\sigma\epsilon\beta}{\gamma^2(n(k)+k_0)^{\alpha_0}}, \end{aligned} \tag{3.8}$$

where we define

$$c := \frac{32L^2}{\gamma^2 n_0} + \frac{24\epsilon\beta L}{\gamma^2\sqrt{n_0}} + \frac{1.1\sigma\epsilon\beta}{\gamma^2 n_0^{1-\alpha_0}} + \left(\epsilon\frac{\beta}{\gamma}\right)^2.$$

We pick n_0 large enough such that there exists $\alpha_0 > 0$ for which $c < 1$.

Unrolling the recursion given by equation (3.8), we get

$$\mathbb{E}\left[\|\theta_{k+1} - \theta_{\mathrm{PS}}\|_2^2\right] \leq c^k\|\theta_1 - \theta_{\mathrm{PS}}\|_2^2 + \frac{1}{\gamma^2}\sum_{j=1}^{k} c^{k-j}\left(\frac{1.2\sigma^2}{n(j)+k_0} + \frac{1.1\sigma\epsilon\beta}{(n(j)+k_0)^{\alpha_0}}\right).$$

Since $\alpha_0 < 1$, we can upper bound the second term as

$$\begin{aligned}&\frac{1}{\gamma^2}\sum_{j=1}^{k} c^{k-j}\left(\frac{1.2\sigma^2}{n(j)+k_0} + \frac{1.1\sigma\epsilon\beta}{(n(j)+k_0)^{\alpha_0}}\right) \\ \leq\ &\frac{1.2\sigma^2}{\gamma^2}\sum_{j=1}^{k} c^{k-j}\frac{1}{n(j)+k_0} + \frac{1.1\sigma\epsilon\beta}{\gamma^2}\sum_{j=1}^{k} c^{k-j}\frac{1}{(n(j)+k_0)^{\alpha_0}} \\ \leq\ &\frac{1}{\gamma^2(1-c)}\left(\frac{1.2\sigma^2}{n_0}(2k^{-\alpha} + c^{(1-2^{-1/\alpha})k}) + \frac{1.1\sigma\epsilon\beta}{n_0^{\alpha_0}}(2k^{-\alpha\cdot\alpha_0} + c^{(1-2^{-1/(\alpha\alpha_0)})k})\right)\end{aligned}$$

where in the second inequality we apply Lemma 3.5.4 after plugging in the choice of $n(k)$. Using the fact that $\alpha_0 \in (0,1)$ and hence $c^{(1-2^{-1/(\alpha\alpha_0)})k} < c^{(1-2^{-1/\alpha})k}$, as well as $\epsilon < \frac{\gamma}{\beta}$ and $n_0 \geq 1$, gives

$$\begin{aligned}
&\frac{1}{\gamma^2(1-c)}\left(\frac{1.2\sigma^2}{n_0}(2k^{-\alpha} + c^{(1-2^{-1/\alpha})k}) + \frac{1.1\sigma\epsilon\beta}{n_0^{\alpha_0}}(2k^{-\alpha\cdot\alpha_0} + c^{(1-2^{-1/(\alpha\alpha_0)})k})\right) \\
\leq\ &\frac{1.2\sigma^2 + 1.1\sigma\gamma}{\gamma^2(1-c)}\left(4k^{-\alpha\alpha_0} + 2c^{(1-2^{-1/\alpha})k}\right) \\
\leq\ &\frac{3(\sigma+\gamma)^2}{\gamma^2(1-c)}\left(2k^{-\alpha\alpha_0} + c^{\Omega(k)}\right).
\end{aligned}$$

It remains to set α_0; we set $\alpha_0 = \max\{\delta \in (0,1) : c < 1\}$ (note that the existence of such α_0 is guaranteed by the choice of n_0). Clearly, $\alpha_0 \to 1$ as n_0 grows, and so putting everything together gives

$$\mathbb{E}\left[\|\theta_{k+1} - \theta_{\mathrm{PS}}\|_2^2\right] \leq c^k\|\theta_1 - \theta_{\mathrm{PS}}\|_2^2 + \frac{3(\sigma+\gamma)^2}{\gamma^2(1-c)}\left(\frac{2}{k^{\alpha\cdot(1-o(1))}} + c^{\Omega(k)}\right),$$

as desired.

Proof of Corollary 3.2.4

From Theorem 3.2.2, we know that for greedy deploy, $\mathbb{E}\left[\|\theta_{k+1} - \theta_{\mathrm{PS}}\|^2\right] = O(\frac{1}{k})$ where k indexes both the number of classifiers and the number of samples collected. By inverting this bound, we see that to ensure $\mathbb{E}\left[\|\theta_{k+1} - \theta_{\mathrm{PS}}\|^2\right] \leq \delta$, it suffices to collect $O(\frac{1}{\delta})$ samples.

From our analogous convergence result for lazy deploy (Theorem 3.2.3), we know that after the k-th deployment, it holds that $\mathbb{E}\left[\|\theta_{k+1} - \theta_{\mathrm{PS}}\|^2\right] = O(1/k^{\alpha\cdot\omega})$, for some $\omega = 1 - o(1)$ which is independent of k and tends to 1 as n_0 grows. If we collect $\Theta(j^\alpha)$ samples for each deployment $j = 1 \ldots k$, after k deployments the total number of samples N is $\Theta(k^{\alpha+1})$. Therefore,

$$\mathbb{E}\left[\|\theta_{k+1} - \theta_{\mathrm{PS}}\|^2\right] = O(1 \ / \ N^{\frac{\alpha\cdot\omega}{\alpha+1}}).$$

By inverting these bounds, we get our desired result for the asymptotics of lazy deploy.

Experimental details

Base distribution. The base distribution consists of the Kaggle data set [40]. We subsample $n = 18,357$ points from the original training set such that both classes are approximately balanced (45% of points have y equal to 1). There are a total of 10 features, 3 of

which we treat as strategic features: utilization of credit lines, number of open credit lines, and number of real estate loans. We scale features in the base distribution so that they have zero mean and unit variance.

Verifying ϵ-sensitivity. We verify that the map $\mathcal{D}(\cdot)$, as described in Section 3.3, is ϵ-sensitive. To do so, we analyze $W_1(\mathcal{D}(\theta), \mathcal{D}(\theta'))$, for arbitrary $\theta, \theta' \in \Theta$. Fix a sample point $x \in \mathbb{R}^{m-1}$ from the base dataset. Because the base distribution $\mathcal{D}$ is supported on n points, we can upper bound the optimal transport distance between any pair of distributions $\mathcal{D}(\theta)$ and $\mathcal{D}(\theta')$ by the Euclidean distance between the shifted versions of x in $\mathcal{D}(\theta)$ and $\mathcal{D}(\theta')$.

In our construction, the point x is shifted to $x - \epsilon\theta$ and to $x - \epsilon\theta'$ in $\mathcal{D}(\theta)$ and $\mathcal{D}(\theta')$ respectively. The distance between these two shifted points is $\|x - \epsilon\theta - x + \epsilon\theta'\|_2 = \epsilon\|\theta - \theta'\|_2$. Since the same relationship holds for all other samples x in the base dataset, the optimal transport from $\mathcal{D}(\theta)$ to $\mathcal{D}(\theta')$ is at most $\epsilon\|\theta - \theta'\|_2$.

Verifying joint smoothness of the objective. For the experiments described in Figure 3.3, we run repeated risk minimization and repeated gradient descent on the logistic loss with ℓ_2 regularization:

$$\frac{1}{n}\sum_{i=1}^{n} -y_i\theta^\top x_i + \log\left(1 + \exp(\theta^\top x_i)\right) + \frac{\gamma}{2}\|\theta\|^2 \tag{3.9}$$

For both the repeated risk minimization and repeated gradient descent we set $\gamma = 1000/n$, where n is the size of the base dataset.

For a particular feature-outcome pair (x_i, y_i), the logistic loss is $\frac{1}{4}\|x_i\|^2$ smooth [76]. Therefore, the entire objective is $\frac{1}{4n}\sum_{i=1}^{n}\|x_i\|^2 + \gamma$ smooth. Due to the strategic updates, $x_{BR} = x - \epsilon\theta$, the norm of individual features change depending on the choice of model parameters.

Theoretically, we can upper bound the smoothness of the objective by finding the implicit constraints on Θ, which can be revealed by looking at the dual of the objective function for every fixed value of ϵ. However, for simplicity, we simply calculate the worst-case smoothness of the objective, given the trajectory of iterates $\{\theta_t\}$, for every fixed ϵ.

Furthermore, we can verify the logistic loss is jointly smooth. For a fixed example $z = (x, y)$, the gradient of the regularized logistic loss with respect to θ is,

$$\nabla_\theta \ell(z;\theta) = yx + \frac{\exp(\theta^\top x)}{1 + \exp(\theta^\top x)}x + \gamma\theta,$$

which is 2-Lipschitz in z due to $y \in \{0, 1\}$. Hence, the overall objective is β-jointly smooth with parameter

$$\beta = \max\left\{2, \frac{1}{4n}\sum_{i=1}^{n} \|x_i\|^2 + \gamma\right\}.$$

For RRM, ϵ is less than $\frac{\gamma}{\beta}$ only in the case that $\epsilon = 0.01$. For RGD, ϵ is never smaller than the theoretical cutoff of $\frac{\gamma}{(\beta+\gamma)(1+1.5\eta\beta)}$.

Optimization details. The definition of RRM requires exact minimization of the objective at every iteration. We approximate this requirement by minimizing the objective described in expression (3.9) to small tolerance, 10^{-8}, using gradient descent. We choose the step size at every iteration using backtracking line search.

In the case of repeated gradient descent, we run the procedure as described in Definition 3.1.4 with a fixed step size of $\eta = \frac{2}{\beta+\gamma}$. For the greedy vs lazy deploy comparison we fix $\lambda = 10^3/n$ for all experiments. When evaluated on the base distribution, the objective has parameters $\beta = 4.72$, $\gamma = 0.054$ which yields $\frac{\gamma}{\beta} = 0.011$.

Chapter 4

In Search of Performative Optima I: Establishing Convexity

Up until this point, the algorithmic results in this thesis have been primarily centered on the possibilities and limitations of finding performatively stable points through retraining.

Recall that stability is a local definition of optimality, by which a model minimizes the expected risk for the specific distribution that it induces. And, *if* performative effects are *very weak*, this local definition of optimality is also nearly globally optimal. In particular, we proved in Theorem 2.2.3 and Corollary 2.2.4 that stable points approximately minimize the performative risk, the central notion of performance in performative prediction.

In full generality however, stability has little bearing on whether a predictive model has low performative risk. If performative effects are not vanishingly small, stable models can in fact *maximize* the performative risk, even for well-behaved, ϵ-sensitive distribution maps with ϵ values that are $O(1)$ (Proposition 2.2.6).

Reasoning by analogy, stable models can be thought of as an *echo chamber* in an online platform. In an echo chamber, one is reassured of their ideas by voicing them, but it's not clear whether they are reasonable outside of this niche community. Similarly, stable solutions minimize risk on the distribution that they induce, but they provide no global guarantees of performance.

In this chapter, we shift attention past performative stability and study optimizing the performative risk directly. Optimizing the performative risk requires a different algorithmic approach than what we've seen so far. Firstly, if performative effects are significant, different predictive models can induce significantly different distributions. Consequently, we need to actively *anticipate* performative effects rather than myopically retrain until convergence, as the latter would only lead to stability. Second, as seen in even in Proposition 2.2.6, the performative risk can be non-convex, even if the loss ℓ is convex

and the distribution map $\mathcal{D}(\cdot)$ has a small Lipschitz constant.[1]. This non-convexity poses computational and statistical barriers when trying to find the performatively optimal solution.

In this chapter, we start to answer the question of when and why can the performative risk be optimized efficiently. In particular:

1. We identify natural conditions under which the performative risk is guaranteed to be *convex*, even in settings where performative effects can be arbitrarily strong.

2. Having established these structural results, we move on to study optimization algorithms which explicitly account for the impacts of prediction and provably and efficiently minimize the performative risk.

4.1 When is the Performative Risk Convex

In this section, we introduce our main structural results illustrating how the performative risk can be convex in various settings, and hence amenable to direct optimization. Throughout our presentation, we adopt the following convention. We state that the performative risk is λ-convex, for some $\lambda \in \mathbb{R}$, if the objective,

$$\mathrm{PR}(\theta) - \frac{\lambda}{2}\|\theta\|_2^2$$

is convex. In other words, if λ is positive, then $\mathrm{PR}(\theta)$ is λ-strongly convex. If λ is negative, then adding the analogous regularizer $\frac{\lambda}{2}\|\theta\|^2$ ensures $\mathrm{PR}(\theta)$ is convex.

To achieve tighter characterization of when the performative risk is convex, we will pay close attention to properties of the loss function with respect to the *data* z. In particular, we will sometimes assume that the loss ℓ is γ_z-strongly convex in z.

That is, for all θ, z, z',

$$\ell(z;\theta) \geq \ell(z';\theta) + \nabla_z \ell(z';\theta)^\top (z' - z) + \frac{\gamma_z}{2}\|z - z'\|_2^2 . \tag{4.1}$$

Assuming convexity in the data is somewhat unorthodox within the stochastic optimization literature since the data is not something we usually optimize over. However, in performative prediction, we do in fact get to actively optimize the data distribution via the dependence distribution map $\mathcal{D}(\cdot)$ on the model parameters θ.[2]

In fact, it is quite natural to assume that we might want to actively *steer* the data towards socially desirable targets $z_\star$ by adding terms like $\|z - z_\star\|_2^2$ to the loss function

[1] In particular, in the counterexample from Proposition 2.2.6, $\mathrm{PR}(\cdot)$ is concave if $\epsilon > \gamma/\beta$

[2] Interestingly, convexity in the data also holds for natural loss functions like squared loss $(y - \theta^\top x)^2$.

ℓ. For examples, in education, we might want induce models that steer people towards achieving certain features x (e.g., higher test scores) or particular outcomes (e.g., increase the likelihood of graduating from high school on time. As we will later see, adding these steering terms to the objective can often make optimization *easier*, not harder.

Lastly, in this chapter we will focus our analysis on problems which satisfy a structural condition we call *mixture dominance*. A distribution map, loss pair $(\mathcal{D}(\cdot), \ell)$ satisfies *mixture dominance* if the following condition holds for all $\theta, \theta', \theta_0 \in \Theta$ and $\alpha \in (0, 1)$:

$$\mathbb{E}_{z\sim\mathcal{D}(\alpha\theta+(1-\alpha)\theta')}\ell(z;\theta_0) \leq \mathbb{E}_{z\sim\alpha\mathcal{D}(\theta)+(1-\alpha)\mathcal{D}(\theta')}\ell(z;\theta_0). \tag{4.2}$$

While previous definitions like smoothness and strong convexity are standard in the optimization literature and have appeared previously in our study of performative prediction, the mixture dominance condition is novel and plays a central role in our analysis of when the performative risk is convex.

To provide some intuition for this condition, we recall the definition of the *decoupled performative risk*:

$$\mathrm{DPR}(\theta, \theta') = \mathbb{E}_{z\sim\mathcal{D}(\theta)}\ell(z;\theta').$$

Notice that asserting convexity of the performative risk is equivalent to showing convexity of $\mathrm{DPR}(\theta, \theta)$ when both arguments are forced to be the same. While convexity of the loss ℓ in θ guarantees that DPR is convex in the second argument, mixture dominance (4.2) essentially posits convexity of DPR in the first argument. Importantly, assuming convexity in each argument separately does *not* directly imply that the performative risk is convex.[3]

Assumption (4.2) is a stochastic dominance statement: the mixture distribution $\alpha\mathcal{D}(\theta)+(1-\alpha)\mathcal{D}(\theta')$ "dominates" $\mathcal{D}(\alpha\theta+(1-\alpha)\theta')$ under a certain loss function. Similar conditions have been extensively studied within the literature on stochastic orders [75]. Part of our analysis relies on incorporating tools from this literature. For example, using results from stochastic orders we can show that (4.2) holds when the loss is convex in z and the distribution map $\mathcal{D}(\cdot)$ forms a *location-scale family* of the form:

$$z_\theta \sim \mathcal{D}(\theta) \;\Leftrightarrow\; z_\theta \overset{d}{=} (\Sigma_0 + \Sigma(\theta))z_{\text{base}} + \mu_0 + \mu\theta, \tag{4.3}$$

where $z_{\text{base}} \sim \mathcal{D}_{\text{base}}$ is a sample from a fixed zero-mean distribution $\mathcal{D}_{\text{base}}$, and $\Sigma(\theta)$, μ are linear maps (see Proposition 4.5.4 for a formal proof).

Location-scale distribution maps of this sort are commonplace throughout the performative prediction literature and hence satisfy mixture dominance if the loss ℓ is convex. For instance, the distribution map for the strategic classification simulations whereby features adapt towards the decision boundary, $x'_S = x_S - \epsilon\theta_S$, is a location family. Other

[3]For example, bilinear objectives $u^\top v$ are convex in (u, v) separately for $u, v \in \mathbb{R}^n$, but not jointly.

examples of location families can be found in previous work on strategic classification [26, 30]. Mixture dominance can also hold in discrete settings, e.g. $\mathcal{D}(\theta) = \text{Bernoulli}(a^\top \theta + b)$ satisfies this condition for any loss. Having provided some context on the mixture dominance condition, we can now state the main result of this section:

Theorem 4.1.1. *Suppose that the loss function $\ell(z; \theta)$ is γ-strongly convex in θ* (2.3)*, β-smooth in z* (3.1)*, and that $\mathcal{D}(\cdot)$ is ϵ-sensitive* (2.2)*. If mixture dominance* (4.2) *holds, then the performative risk is λ-convex for $\lambda = \gamma - 2\epsilon\beta$.*

One interesting facet of this result is that it shows that under the mixture dominance condition, $\epsilon = \gamma/(2\beta)$ is a sharp threshold for $\text{PR}(\cdot)$ to be convex in θ. In particular, if $\epsilon < \gamma/(2\beta)$, the theorem above guarantees that the performative risk if guaranteed to be strongly convex. Furthermore, in the proof of Proposition 2.2.6, we saw that there exists a performative prediction problem satisfying the mixture dominance condition that is β-smooth in z, γ-strongly convex in θ, and ϵ-sensitive for which $\text{PR}(\cdot)$ is *concave* in θ if $\epsilon > \gamma/(2\beta)$. Hence, $\gamma/(2\beta)$ is a sharp threshold for convexity of the performative risk.

Remark 4.1.2. Notice that this is the second time we find the existence of sharp thresholds in performative prediction. Recall from Chapter 3 that $\epsilon = \gamma/\beta$ is a sharp threshold for the convergence of repeated retraining (Definition 2.1.6). Now, we see that dividing this quantity by 2, we get another threshold for the convexity of the performative risk.

Summarizing, assume that mixture-dominance condition holds, that the loss is γ-strongly convex in θ, β-smooth in z, and that the distribution map $\mathcal{D}$ is ϵ-sensitive:

- If $\epsilon > \gamma/\beta$, performative stable point may not exist and retraining may not converge.
- If $\epsilon \in (\gamma/(2\beta), \gamma/\beta)$, retraining is guaranteed to converge to a unique stable point, but this stable point may maximize the performative risk. Furthermore, the performative risk $\text{PR}(\cdot)$ may be non-convex.
- If $\epsilon < \gamma/(2\beta)$, repeated retraining converges to stability, but performative optimality is also within reach since $\text{PR}(\theta)$ is now guaranteed to be convex.

While the threshold $\epsilon = \gamma/(2\beta)$ is in general tight as argued above, for certain families of distribution maps the conclusion of Theorem 4.1.1 can be made considerably stronger. Indeed, in some cases the performative risk is convex *regardless* of the magnitude of performative effects, as observed in the following example.

Example 2.1.1 (continued). Recall the financial trading example where the true prices of financial instruments are nudged to be closer towards our predictions $\hat{y} = \theta^\top x_{\text{base}}$ by

some constant ϵ',

$$(x, y) \sim \mathcal{D}(\theta) \iff (x_{\text{base}}, y_{\text{base}}) \sim \mathcal{D}_{\text{base}} \text{ and } (x, y) = \left(x_{\text{base}},\ y_{\text{base}} + \epsilon' \cdot (\theta^\top x_{\text{base}} - y_{\text{base}})\right)$$

If ℓ is the squared loss $(y - \theta^\top x)^2$, then the performative risk is always convex regardless of the magnitude of ϵ'. In particular, a short calculation shows that:

$$\text{PR}(\theta) = (1 - \epsilon')^2 \mathbb{E}_{(x_{\text{base}}, y_{\text{base}}) \sim \mathcal{D}_{\text{base}}} (y_{\text{base}} - \theta^\top x_{\text{base}})^2$$

Motivated by this observation, we specialize the analysis in Theorem 4.1.1 to the particular case of location-scale families, and obtain a result that is at least as tight as the previous theorem.

Theorem 4.1.3. *Suppose that $\ell(z; \theta)$ is γ-strongly convex in θ (2.3), β-smooth in z (3.1), and γ_z-strongly convex in z (4.1). Furthermore, suppose that $\mathcal{D}(\theta)$ forms a location-scale family (4.3) with ϵ as its sensitivity parameter[4]. Define $\Sigma_{z_{\text{base}}}$ to be the covariance matrix of $z_{\text{base}} \sim \mathcal{D}_0$, and let*

$$\sigma_{\min}(\mu) = \min_{\|\theta\|_2 = 1} \|\mu\theta\|_2, \sigma_{\min}(\Sigma) = \min_{\|\theta\|_2 = 1} \|\Sigma_{z_{\text{base}}}^{1/2} \Sigma(\theta)^\top\|_F.$$

Then, the performative risk is λ-convex for λ equal to:

$$\max\{\gamma - \beta^2/\gamma_z, \gamma - 2\epsilon\beta + \gamma_z(\sigma_{\min}^2(\mu) + \sigma_{\min}^2(\Sigma))\}.$$

The key difference relative to the previous theorem, is that now we take into account the fact that the loss function can be strongly convex in the data z. This property can make convexity easier to establish as we will now illustrate in Example 4.1.4. Under the location-scale assumption, we can add regularizers of the form $\alpha/2\|z - z_\star\|_2^2$ to ensure convexity of the performative risk. These quadratic terms are α-convex in z, meaning that if the previous objective was γ_z convex, it is now $\gamma_z + \alpha$ convex. Note that the smoothness parameter (with respect to z) remains unchanged since a loss is β_z smooth if the gradient of ℓ *with respect to* θ is β_z-Lipschitz. Since $\nabla_\theta(\alpha/2\|z - z_\star\|_2^2) = 0$ this constant remains unchanged!

In general, one can achieve a tighter analysis of when the performative risk is convex by distinguishing between variables which are *static*, whose distribution is the same under $\mathcal{D}(\theta)$ for all θ, and performative variables which are influenced by the deployed classifier. For the most part we avoid this distinction in the main body of this chapter for the sake of readability, however, we elaborate on how the analysis can be strengthened in the supplementary material for this chapter. We now illustrate an application of Theorem 4.1.3 on a scale family example.

[4]The sensitivity parameter ϵ for location-scale families can be explicitly bounded in terms of the parameters μ and $\Sigma(\theta)$; see the supplementary material of this chapter.

Example 4.1.4. Suppose that $x > 0$ is a one-dimensional feature drawn from a fixed distribution $\mathcal{D}_x$, and let $y|x \sim \theta x \cdot \mathrm{Exp}(1)$ be distributed as an exponential random variable with mean θx. Let the loss be the squared loss, $\ell((x, y); \theta) = \frac{1}{2}(y - \theta \cdot x)^2$ and let $\Theta = \mathbb{R}_+$. Note that this example exhibits a self-fulfilling prophecy property whereby all solutions are performatively stable. On the other hand, $\mathrm{PR}(\theta) = \theta^2 \mathbb{E}x^2$, and the unique performative optimum is $\theta_{\mathrm{PO}} = 0$. Again, we see how stability has no bearing on whether a solution has low performative risk.

However, we note that the loss is 1-strongly convex in y. Furthermore, by averaging over the static features, we observe that $\mathrm{PR}(\theta)$ is $\mathbb{E}x^2$-strongly convex in θ and $\mathbb{E}x$-smooth in y. Therefore, according to Theorem 4.1.3, the performative risk is convex and hence tractable to optimize, since $\gamma - \beta^2/\gamma_z = \mathbb{E}x^2 - (\mathbb{E}x)^2 \geq 0$ by Jensen's inequality.

While this example, like most others in this section, is intended as a toy problem to provide the reader with some intuition regarding the intricacies of performativity, many instances of performative prediction in the real world do exhibit a self-fulfilling prophecy aspect whereby predicting a particular outcome increases the likelihood that it occurs. For instance, predicting that a student is unlikely to do well on a standardized exam may discourage them from studying in the first place and hence lower their final grade. Settings like these where stability is a vacuous guarantee of performance remind us how developing reliable predictive models requires going outside the stability echo chamber.

As a final note, to prove the results in this section, we have imposed additional assumptions such as mixture dominance, or analyzed the special case of location-scale families. The reader might naturally ask whether these settings are so restrictive that one can optimize the performative risk using previous optimization methods for performative prediction which find stable points. Or in particular,

If the performative risk is convex,
are performative optima and performative stable points now the same?

It turns out that both solutions can still have qualitatively different behavior, regardless of the strength of performative effects or the convexity of $\mathrm{PR}(\cdot)$. First, notice that the example in the proof of Proposition 2.2.6 is a location family, and as such it satisfies mixture dominance. In that example, when $\epsilon \in (\frac{\gamma}{2\beta}, \frac{\gamma}{\beta})$, methods for finding stable points converge to a maximizer of the performative risk; however, this is outside the regime where the performative risk is convex. In what follows, by relying on Theorem 4.1.3, we provide another scale family example where the performative risk is convex regardless of ϵ, yet stable points can be arbitrarily suboptimal.

Example 4.1.5. Suppose that $\mathcal{D}(\theta) = \mathcal{N}(\mu, \epsilon^2\theta^2)$ for some $\mu \in \mathbb{R}$ and $\epsilon > 0$. This distribution map is ϵ-sensitive. Furthermore, if ℓ is the squared loss, $\ell(z;\theta) = \frac{1}{2}(z-\theta)^2$, then there is a unique stable point $\theta_{\mathrm{PS}} = \mu$. On the other hand, $\theta_{\mathrm{PO}} = \mu/(1+\epsilon^2)$.

Notice how, contrary to the performative optimum θ_{PO}, the stable point θ_{PS} is independent of ϵ and hence oblivious to the performative effects. Depending on μ, the stable point can be arbitrarily suboptimal, since $\mathrm{PR}(\theta_{\mathrm{PS}}) - \mathrm{PR}(\theta_{\mathrm{PO}}) = \Omega(\mu^2)$. Note also that, according to Theorem 4.1.3, the performative risk is $\gamma - 2\epsilon\beta + \gamma_z\sigma^2_{\min}(\Sigma) = 1 - 2\epsilon + \epsilon^2$-convex. Since $1 - 2\epsilon + \epsilon^2 = (\epsilon - 1)^2 \geq 0$, the performative risk is always convex and hence tractable to optimize.

4.2 Optimization Algorithms

Having identified conditions under which the performative risk is convex, we now consider methods for efficiently optimizing it. One of the main challenges of carrying out this task is that, even in convex settings, the learner can only access the objective via noisy function evaluations corresponding to model deployments. Without knowledge of the underlying distribution map $\mathcal{D}(\cdot)$, it is infeasible to (exactly) compute gradients of the performative risk. A naive solution is to apply a zeroth-order method, however, these algorithms are in general hard to tune, and their performance scales poorly with the problem dimension.

Our main algorithmic contribution is to show how one can address these issues by creating an explicit *model* of the distribution map and then optimizing a proxy objective for the performative risk offline. We refer to this as the two-stage procedure for optimizing the performative risk and show it is provably efficient for the case of location families.

To develop further intuition, consider the following simple example. Let $z \sim \mathcal{N}(\epsilon\theta, 1)$ be a one-dimensional Gaussian and let $\ell(z;\theta) = \frac{1}{2}(z-\theta)^2$ be the squared loss. Then, the performative risk, $\mathrm{PR}(\theta) = \frac{1}{2}(\epsilon-1)^2\theta^2$, is a simple, convex function for all values of ϵ (as indeed confirmed by Theorem 4.1.3, since $\gamma - 2\epsilon\beta + \gamma_z\sigma^2_{\min}(\mu) = 1 - 2\epsilon + \epsilon^2 \geq 0$). However, gradients are unavailable since they depend on the density of $\mathcal{D}(\theta)$, denoted p_θ, which is typically unknown:

$$\begin{aligned}\nabla_\theta \mathrm{PR}(\theta) &= \mathbb{E}_{z\sim\mathcal{D}(\theta)}\nabla_\theta\ell(z;\theta) + \mathbb{E}_{z\sim\mathcal{D}(\theta)}\ell(z;\theta)\nabla_\theta \log p_\theta(z) \\ &= \mathbb{E}_{z\sim\mathcal{D}(\theta)} -(z-\theta) + \epsilon(\epsilon-1)\theta.\end{aligned}$$

Despite the simplicity of this example, the earlier approaches to optimization in performative prediction, such as repeated retraining, fail on this problem. The reason is that they essentially ignore the second term in the gradient computation which requires explicitly anticipating performative effects. For example, retraining computes the sequence

of updates

$$\theta_{t+1} = \arg\min_{\theta} \mathbb{E}_{z\sim\mathcal{D}(\theta_t)} \tfrac{1}{2}(z-\theta)^2 = \epsilon\theta_t,$$

which diverges for $|\epsilon| > 1$.

Generic Derivative-Free Methods

Having observed the difficulty of computing gradients, the most natural starting point for optimizing the performative risk is to consider derivative-free methods for convex optimization [2, 25, 78]. These methods work by constructing a noisy estimate of the gradient by querying the objective function at a randomly perturbed point around the current iterate.

For instance, Flaxman et al. [25] sample a vector $u \sim \mathrm{Unif}(\mathcal{S}^{d-1})$ to get a slightly biased gradient estimator,[5]

$$\nabla_\theta \mathrm{PR}(\theta) \approx \frac{d}{\delta}\mathbb{E}[\mathrm{PR}(\theta+\delta u)u],$$

for some small $\delta > 0$. Generic derivative-free algorithms for convex optimization require few assumptions beyond those given in the previous section to ensure convexity. Moreover, they guarantee convergence to a performative optimum given sufficiently many samples. Unfortunately, their rate of convergence can be slow and scales poorly with the problem dimension. In general, zeroth-order methods require $\tilde{O}(d^2/\Delta^2)$ samples to obtain a Δ-suboptimal point [2, 78], which can be prohibitively expensive if samples are hard to come by.

Two-Stage Approach

In cases where we have further structure, an alternative solution to derivative-free methods is to utilize a *two-stage* approach to optimizing the performative risk. In the first stage, we estimate a coarse model of the distribution map, $\widehat{\mathcal{D}}(\cdot)$ via experiment design. Then, in the second stage, the algorithm optimizes a proxy to the performative risk treating the estimated $\widehat{\mathcal{D}}$ as if it were the true distribution map:

$$\hat{\theta}_{\mathrm{PO}} \in \arg\min_{\theta} \widehat{\mathrm{PR}}(\theta) := \mathbb{E}_{z\sim\widehat{\mathcal{D}}(\theta)} \ell(z;\theta).$$

The exact implementation of this idea depends on the problem setting at hand; to make things concrete, we instantiate the approach in the context of location families and prove

[5] We use $\mathcal{S}^{d-1}$ to denote the unit sphere in $\mathbb{R}^d$

Stage 1: Construct a model of the distribution map
(Estimate location parameter μ with experiment design)

For $i = 1, \dots n$:

- Sample and deploy classifier $\theta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$.
- Observe $z_i \sim \mathcal{D}(\theta_i)$.

Estimate μ via least squares:

$$\hat{\mu} \in \arg\min_{\mu} \sum_{i=1}^{n} \|z_i - \mu\theta_i\|_2^2$$

Gather samples from the base distribution

For $j = n + 1, \dots 2n$:

- Deploy classifier $\theta_j = 0$, and observe $z_j \sim \mathcal{D}(0)$.

Stage 2: Minimize a finite-sample approximation of the performative risk,

$$\arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{j=n+1}^{2n} \ell(z_j + \hat{\mu}\theta; \theta).$$

Figure 4.1: Two-Stage Algorithm for Location Families.

that it optimizes the performative risk with significantly better sample complexity than generic zeroth-order methods. For the remainder of this section, we assume the distribution map $\mathcal{D}$ is parameterized by a location family

$$z_\theta \sim \mathcal{D}(\theta) \;\Leftrightarrow\; z_\theta \overset{d}{=} z_{\text{base}} + \mu\theta,$$

where the matrix $\mu \in \mathbb{R}^{m \times d}$ is an unknown parameter, and $z_{\text{base}} \sim \mathcal{D}_0$ is a zero-mean random variable.[6]

In the first stage of our two-stage procedure we build a model of the distribution map $\widehat{\mathcal{D}}$ that in effect allows us to draw samples $z \sim \widehat{\mathcal{D}}(\theta) \approx \mathcal{D}(\theta)$. To do this, we

[6]The variable z_0 being zero-mean is only to simplify the exposition; the same analysis carries over when there is an additional intercept term. Similarly, the choice of Gaussian noise in the experiment design phase of Algorithm 4.1 is made for convenience. In general, any subgaussian distribution with full rank covariance would suffice.

perform experiment design to recover the unknown parameter μ which captures the performative effects. In particular, we sample and deploy n classifiers θ_i, $i \in [n]$, observe data $z_i \sim \mathcal{D}(\theta_i)$, and then construct an estimate $\hat{\mu}$ of the location map μ using ordinary least squares. We then gather samples from the base distribution $\mathcal{D}_0$ by repeatedly deploying the zero classifier. In the location-family model, deploying the zero classifier ensures we observe data points z_{base}, without performative effects. With both of these components, given any θ', we can simulate $z \sim \widehat{\mathcal{D}}(\theta')$ by taking $z = z_{\text{base}} + \hat{\mu}\theta'$.

In the second stage, we use the estimated model to construct a proxy objective. Define the perturbed performative risk:

$$\widehat{\text{PR}}(\theta) = \mathbb{E}_{z \sim \widehat{\mathcal{D}}(\theta)} \ell(z; \theta) = \mathbb{E}_{z_{\text{base}} \sim \mathcal{D}_0} \ell(z_{\text{base}} + \hat{\mu}\theta; \theta).$$

Note that $\text{PR}(\theta) = \mathbb{E}_{z_{\text{base}} \sim \mathcal{D}_0} \ell(z_{\text{base}} + \mu\theta; \theta)$. Using the estimated parameter $\hat{\mu}$ and samples $z_i \sim \mathcal{D}_0$, we can construct a finite-sample approximation to the perturbed performative risk and find the following optimizer:

$$\hat{\theta}_n \in \arg\min_{\theta \in \Theta} \widehat{\text{PR}}_n(\theta) := \frac{1}{n} \sum_{i=n+1}^{2n} \ell(z_i + \hat{\mu}\theta; \theta).$$

The main technical result in this section shows that, under appropriate regularity assumptions on the loss, Algorithm 4.1 efficiently approximates the performative optimum. In particular, when the data dimensionality m is comparable to the model dimensionality d, i.e. $m = O(d)$, then computing a Δ-suboptimal classifier requires $O(d/\Delta)$ samples. In contrast, the derivative-free methods considered previously require $\widetilde{O}(d^2/\Delta^2)$ samples to compute a classifier of similar quality. The formal statement and proof of this result is deferred to the supplementary material for this chapter.

Theorem 4.2.1 (Informal)**.** *Under appropriate smoothness and strong convexity assumptions on the loss ℓ, if the distribution of z_{base} is sub-Gaussian, and if the number of samples $n \geq \Omega(d + m + \log(1/\delta))$, then, with probability $1 - \delta$, Algorithm 4.1 returns a point $\hat{\theta}_n$ such that*

$$\text{PR}(\hat{\theta}_n) - \text{PR}(\theta_{\text{PO}}) \leq O\left(\frac{d + m + \log(1/\delta)}{n} + \frac{1}{\delta n}\right).$$

While we analyze this two-stage procedure in the context of location families, the principles behind the approach can be extended to more general settings. Whenever the distribution map has enough structure to efficiently estimate a model $\widehat{\mathcal{D}}$ that supports sampling new data, we can always use the "plug-in" approach above and construct and optimize a perturbed version of the performative risk.

4.3 Simulations

We complement our theoretical analysis with empirical evaluations on the semi-synthetic strategic classification simulation introduced previously in Chapter 3.

In our experiments, we pay particular attention to understanding the differences in empirical performance between algorithms which converge to performative optima, such as the two-stage procedure or derivative-free methods, versus the optimization algorithms which converge to performatively stable points, such as greedy and lazy SGD.

In addition, we examine the differences in the sample efficiency of the different algorithms and examine their sensitivity to the relevant structural assumptions outlined in Section 4.1. To evaluate derivative-free methods, we implement the "gradient descent without a gradient" algorithm from [25], which we refer to from here on out as the "DFO algorithm." For each of the following experiments, we run each algorithm 50 times and display 95% bootstrap confidence intervals. We provide a formal description of all the procedures, as well as a detailed description of the experimental setup in Section 4.5.

We briefly review the details regarding the strategic classification simulator from Chapter 3. The simulator models a strategic classification problem between a bank and individual agents seeking a loan. The bank deploys a logistic regression classifier f_θ to determine the individuals' default probabilities, while agents strategically manipulate their features to achieve a more favorable classification. The goal of the bank is to find the performative optimal model, or Stackelberg equilibria that minimizes the performative risk.

As detailed previously, the distribution map for this problem is a location family and is ϵ-sensitive. The institution's loss is an ℓ_2 regularized logistic regression objective.

Since the logistic loss is not strongly convex in the features, we only have a certificate of convexity when ϵ is small enough (namely, $\epsilon \leq \frac{\gamma}{2\beta}$). We consider two values of ϵ: one which is below this critical threshold ($\epsilon = .0001$), and one large value for which we do not have theoretical guarantees ($\epsilon = 100$). When ϵ is small, both the DFO algorithm and the two-stage method yield significantly higher accuracy solutions compared to the two variants of SGD (see left panel of Figure 4.2). This observation serves as further evidence that stable points have significantly worse performative risk relative to performative optima, even in regimes where $\epsilon < \gamma/(2\beta)$.

Also note that, although both the DFO algorithm and the two-stage algorithm improve upon methods for repeated retraining, the two-stage algorithm converges with significantly fewer samples and significantly lower variance. Indeed, a few thousand samples suffice for convergence of the two-stage method, whereas the DFO algorithm has still not fully converged after a million samples.

Lastly, on the top right plot, we evaluate these methods for $\epsilon \gg \gamma/(2\beta)$ which is outside the regime of our theoretical analysis. Consequently, we have no convergence guarantees

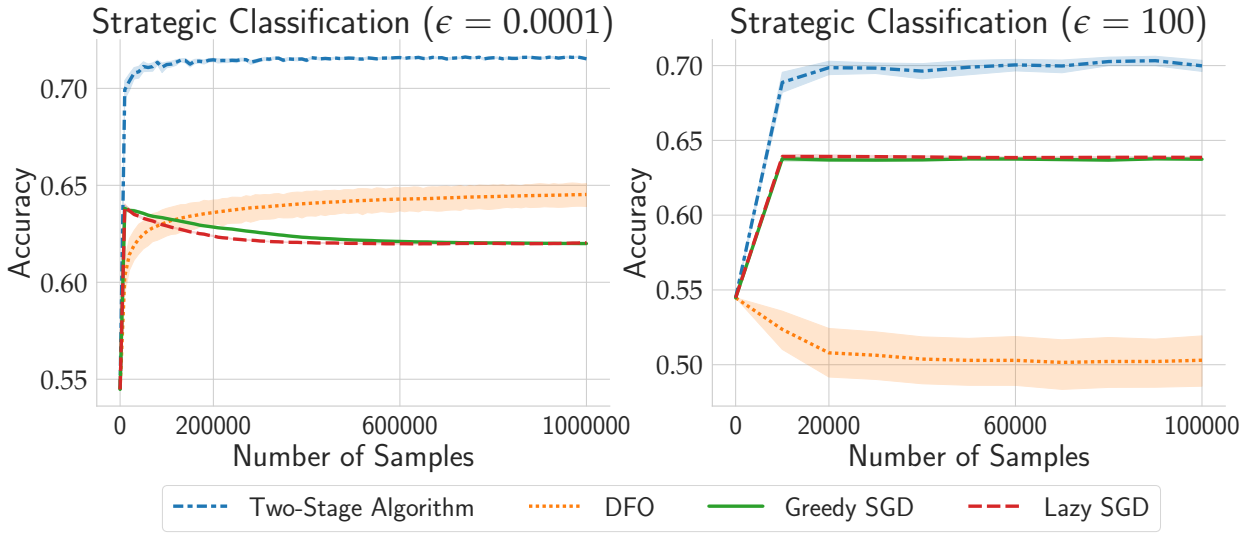


Figure 4.2: Classification accuracy versus number of samples collected for the two-stage algorithm, DFO algorithm, greedy SGD, and lazy SGD, for $\epsilon = 0.0001 \leq \frac{\gamma}{2\beta}$ (left) and $\epsilon = 100 \gg \frac{\gamma}{2\beta}$ (right). Each experiment is repeated 50 times, and we display 95% bootstrap confidence intervals.

for any of the four algorithms. Despite the lack of guarantees and the increased strength of performative effects, we see that the two-stage procedure achieves only a slightly lower accuracy than in the previous setting. On the other hand, as described in our echo chamber analogy, greedy and lazy SGD rapidly converge to a local minimum and do not significantly improve predictive performance after the 10k sample mark. Despite extensive tuning, we were unable to improve the performance of the DFO algorithm and achieve nontrivial accuracy with this method.

4.4 Chapter Notes

The results from this chapter were first published in [59]. In addition to this work, there have been a number of other papers studying the possibility of directly optimizing the performative risk. [36] directly assumes convexity of the performative risk and proposes solving for performative optima by zeroth-order gradient descent. This algorithm is later extended to work in settings where the distribution is stateful and gradually adapts to the latest deployed classifier [38].

In a different line of work, [39] study bandit style online algorithms for optimizing the performative risk in cases where the objective need not be convex. They illustrate how performative prediction exhibits a distinctly richer kind of feedback that allows for faster rates of optimization. [64] study analogous approaches towards finding performatively optimal solutions in the multiplayer performative prediction setting. Lastly, [90] study the possibilities of finding performative optima in places where the performative risk is only weakly convex.

4.5 Supplementary Material

Background on Stochastic Orders

In this section we provide the necessary preliminaries from the literature on stochastic orders.

First, we recall the notion of the *convex order*: for two random vectors $z, z' \in \mathbb{R}^m$, we say that z is less than z' in the convex order, denoted $z \preceq_{cx} z'$, if for all convex functions $g : \mathbb{R}^m \to \mathbb{R}$, it holds that

$$\mathbb{E}g(z) \leq \mathbb{E}g(z').$$

Using a slight abuse of notation, we will also write $\mathcal{D}_1 \preceq_{cx} \mathcal{D}_2$ for two distributions $\mathcal{D}_1, \mathcal{D}_2$ when $z \sim \mathcal{D}_1, z' \sim \mathcal{D}_2$ and $z \preceq_{cx} z'$.

Therefore, an immediate way to satisfy condition (4.2) is to assume that the loss function $\ell(z;\theta)$ is convex in z, and to require $\mathcal{D}(\alpha\theta + (1-\alpha)\theta') \preceq_{cx} \alpha\mathcal{D}(\theta) + (1-\alpha)\mathcal{D}(\theta')$. The latter condition has been long studied in classical statistical literature and many equivalent characterizations are known (see, e.g., [63, 72, 75]). This leads to the following corollary of Theorem 4.1.1.

Corollary 4.5.1. *Suppose that the loss function is γ-strongly convex in θ* (2.3) *and β-smooth in z* (3.1)*, and that the distribution map $\mathcal{D}(\cdot)$ is ϵ-sensitive* (2.2)*. Further, assume that $\ell(z;\theta)$ is convex in z* (4.1) *and that $\mathcal{D}(\alpha\theta + (1-\alpha)\theta') \preceq_{cx} \alpha\mathcal{D}(\theta) + (1-\alpha)\mathcal{D}(\theta')$. Then, the performative risk* $\mathrm{PR}(\theta)$ *is $(\gamma - 2\epsilon\beta)$-convex.*

Now we discuss important families of distributions that satisfy the convex order condition $\mathcal{D}(\alpha\theta + (1-\alpha)\theta') \preceq_{cx} \alpha\mathcal{D}(\theta) + (1-\alpha)\mathcal{D}(\theta')$.

Example 4.5.2. An obvious example where $\mathcal{D}\left(\alpha\theta + (1-\alpha)\theta'\right) \preceq_{cx} \alpha\mathcal{D}(\theta) + (1-\alpha)\mathcal{D}(\theta')$ is when $\mathcal{D}(\alpha\theta + (1-\alpha)\theta') = \alpha\mathcal{D}(\theta) + (1-\alpha)\mathcal{D}(\theta')$. An important setting which satisfies this linearity property is when the probability of a positive outcome of a binary variable is linear in θ: $z_\theta \sim \mathrm{Bern}\left(a + w^\top\theta\right)$ defines $z_\theta \sim \mathcal{D}(\theta)$. In this case, $\mathcal{D}(\alpha\theta + (1-\alpha)\theta') = \alpha\mathcal{D}(\theta) + (1-\alpha)\mathcal{D}(\theta')$.

For further examples, we invoke a convenient characterization of the convex order condition.

Lemma 4.5.3 ([62])**.** *Two random vectors z and z' satisfy $z \leq_{cx} z'$ if and only if there exists a coupling of z and z' such that $\mathbb{E}[z'|z] = z$ a.s.*

By applying Lemma 4.5.3, we show that the important case of *location-scale families* satisfies the convex order condition. Therefore, if the loss function is additionally convex in z, condition (4.2) follows.

Proposition 4.5.4. *Suppose that $\mathcal{D}(\theta)$ forms a location-scale family* (4.3) *such that $\Sigma_0 + \Sigma(\theta)$ has full rank for all $\theta \in \Theta$. Then, $\mathcal{D}(\alpha\theta + (1-\alpha)\theta') \leq_{cx} \alpha\mathcal{D}(\theta) + (1-\alpha)\mathcal{D}(\theta')$ for all $\theta, \theta' \in \Theta$.*

Proof. We will construct a coupling (z, z') such that $z \sim \mathcal{D}(\alpha\theta + (1-\alpha)\theta')$, $z' \sim \alpha\mathcal{D}(\theta) + (1-\alpha)\mathcal{D}(\theta')$, and $\mathbb{E}[z'|z] = z$. Let $z \sim \mathcal{D}(\alpha\theta + (1-\alpha)\theta')$; then we define z' in terms of z as

$$z' = (\Sigma_0 + \Sigma(G))(\Sigma_0 + \Sigma(\alpha\theta + (1-\alpha)\theta'))^{-1}(z - \mu_0 - \mu(\alpha\theta + (1-\alpha)\theta')) + \mu_0 + \mu G, \quad (4.4)$$

where

$$G = \begin{cases} \theta, \text{ with probability } \alpha, \\ \theta', \text{ with probability } 1-\alpha \end{cases}$$

is independent of z. Notice that $\mathbb{E}[z' \mid z]$ is equal to

$$\begin{aligned} &= \mathbb{E}\left[(\Sigma_0 + \Sigma(G))(\Sigma_0 + \Sigma(\alpha\theta + (1-\alpha)\theta'))^{-1}(z - \mu_0 - \mu(\alpha\theta + (1-\alpha)\theta')) + \mu_0 + \mu G \mid z\right] \\ &= (\Sigma_0 + \mathbb{E}[\Sigma(G)])(\Sigma_0 + \Sigma(\alpha\theta + (1-\alpha)\theta'))^{-1}(z - \mu_0 - \mu(\alpha\theta + (1-\alpha)\theta')) + \mu_0 + \mathbb{E}[\mu G] \\ &= z, \end{aligned}$$

which follows by linearity of μ and $\Sigma(\cdot)$ and the fact that $\mathbb{E}[G] = \alpha\theta + (1-\alpha)\theta'$.

We now only need to verify that $z' \sim \alpha\mathcal{D}(\theta) + (1-\alpha)\mathcal{D}(\theta')$ in order to apply Lemma 4.5.3 and conclude that $z' \leq_{cx} z$. Indeed, with probability α we have $G = \theta$, and on that event $z' \overset{d}{=} (\Sigma_0 + \Sigma(\theta))z_{\text{base}} + \mu_0 + \mu\theta$; a similar argument applies to θ'. Therefore, putting everything together we conclude that $z \leq_{cx} z'$. □

Proposition 4.5.4 implies that for all convex functions $g : \mathbb{R}^m \to \mathbb{R}$,

$$\mathbb{E}_{z\sim\mathcal{D}(\alpha\theta+(1-\alpha)\theta')}[g(z)] \leq \mathbb{E}_{z\sim\alpha\mathcal{D}(\theta)+(1-\alpha)\mathcal{D}(\theta')}[g(z)].$$

We now show that for *strongly* convex g, this conclusion can be made even stronger.

Proposition 4.5.5. *Let $g : \mathbb{R}^m \to \mathbb{R}$ be a γ_z-strongly convex function* (4.1) *for some $\gamma_z \geq 0$, and let $\mathcal{D}(\theta)$ form a location-scale family* (4.3)*. Then,*

$$\mathbb{E}_{z\sim\mathcal{D}(\alpha\theta+(1-\alpha)\theta')}[g(z)] \leq \mathbb{E}_{z\sim\alpha\mathcal{D}(\theta)+(1-\alpha)\mathcal{D}(\theta')}[g(z)] - \frac{\alpha(1-\alpha)\gamma_z}{2}\mathbb{E}\|\Sigma(\theta-\theta')z_{\text{base}} + \mu(\theta-\theta')\|_2^2.$$

Proof. Since g is strongly convex, we can write $g(z) = g_0(z) + \frac{\gamma_z}{2}\|z\|_2^2$, where g_0 is a convex function. Thus, we want to prove

$$\mathbb{E}_{z\sim\mathcal{D}(\alpha\theta+(1-\alpha)\theta')}\left[g_0(z) + \frac{\gamma_z}{2}\|z\|_2^2\right] \leq \mathbb{E}_{z'\sim\alpha\mathcal{D}(\theta)+(1-\alpha)\mathcal{D}(\theta')}\left[g_0(z') + \frac{\gamma_z}{2}\|z'\|_2^2\right] - \frac{\alpha(1-\alpha)\gamma_z}{2}\mathbb{E}\left\|\Sigma(\theta-\theta')z_{\text{base}} + \mu(\theta-\theta')\right\|_2^2.$$

By Proposition 4.5.4, we know that

$$\mathbb{E}_{z\sim\mathcal{D}(\alpha\theta+(1-\alpha)\theta')}[g_0(z)] \leq \mathbb{E}_{z\sim\alpha\mathcal{D}(\theta)+(1-\alpha)\mathcal{D}(\theta')}[g_0(z)].$$

Therefore, we only need to argue that

$$\mathbb{E}\left[\|z'\|_2^2 - \|z\|_2^2\right] \geq \alpha(1-\alpha)\mathbb{E}\|\Sigma(\theta-\theta')z_{\text{base}} + \mu(\theta-\theta')\|_2^2.$$

Without loss of generality, we take z, z' to be coupled as in equation (4.4). Then, we can write

$$\begin{aligned}\mathbb{E}\left[\|z'\|_2^2 - \|z\|_2^2\right] &= \mathbb{E}\left[\|z'-z\|_2^2 + 2(z'-z)^\top z\right] \\ &= \mathbb{E}\left[\|z'-z\|_2^2\right] \\ &= \mathbb{E}\left[\left\|\Sigma\left(G - (\alpha\theta + (1-\alpha)\theta')\right)z_{\text{base}} + \mu\left(G - (\alpha\theta+(1-\alpha)\theta')\right)\right\|_2^2\right],\end{aligned}$$

where the second steps follows by iterating expectations, because $\mathbb{E}[z'|z] = z$.

By further taking an expectation over G, we get:

$$\begin{aligned}&\mathbb{E}\left[\left\|\Sigma\left(G - (\alpha\theta + (1-\alpha)\theta')\right)z_{\text{base}} + \mu\left(G - (\alpha\theta+(1-\alpha)\theta')\right)\right\|_2^2\right] \\ &= \alpha(1-\alpha)^2\mathbb{E}\|\Sigma(\theta-\theta')z_{\text{base}} + \mu(\theta-\theta')\|_2^2 + (1-\alpha)\alpha^2\mathbb{E}\|\Sigma(\theta'-\theta)z_{\text{base}} + \mu(\theta-\theta')\|_2^2 \\ &= \alpha(1-\alpha)\mathbb{E}\|\Sigma(\theta-\theta')z_{\text{base}} + \mu(\theta-\theta')\|_2^2.\end{aligned}$$

□

Proving Convexity of the Performative Risk

Proof of Theorem 4.1.1

We begin by writing out the gradient of the performative risk:

$$\begin{aligned}\nabla_\theta \text{PR}(\theta) = \nabla_\theta\left(\int \ell(z;\theta)p_\theta(z)dz\right) &= \int \nabla_\theta\ell(z;\theta)p_\theta(z)dz + \int \ell(z;\theta)\nabla_\theta p_\theta(z)dz \\ &= \int \nabla_\theta\ell(z;\theta)p_\theta(z)dz + \int \ell(z;\theta)\nabla_\theta\log(p_\theta(z))p_\theta(z)dz \\ &= \mathbb{E}_{z\sim\mathcal{D}(\theta)}[\nabla_\theta\ell(z;\theta)] + \mathbb{E}_{z\sim\mathcal{D}(\theta)}[\ell(z;\theta)\nabla_\theta\log(p_\theta(z))].\end{aligned}$$

By the first-order condition for convexity, we know that $\mathrm{PR}(\theta)$ is $(\gamma - 2\epsilon\beta)$-convex if and only if

$$\left(\mathbb{E}_{z\sim\mathcal{D}(\theta)}[\nabla_\theta \ell(z;\theta) + \ell(z;\theta)\nabla_\theta \log(p_\theta(z))]\right)^\top (\theta' - \theta) + \frac{\gamma - 2\epsilon\beta}{2}\|\theta - \theta'\|_2^2 \leq \mathrm{PR}(\theta') - \mathrm{PR}(\theta), \tag{4.5}$$

for all $\theta, \theta' \in \Theta$. By assumption (4.2), we know that for all $\theta, \theta', \theta_0 \in \Theta$,

$$\mathbb{E}_{z\sim\mathcal{D}(\alpha\theta+(1-\alpha)\theta')}[\ell(z;\theta_0)] \leq \alpha\mathbb{E}_{z\sim\mathcal{D}(\theta)}[\ell(z;\theta_0)] + (1-\alpha)\mathbb{E}_{z\sim\mathcal{D}(\theta')}[\ell(z;\theta_0)].$$

This assumption is equivalent to saying that $g_{\theta_0}(\theta) = \mathbb{E}_{z\sim\mathcal{D}(\theta)}[\ell(z;\theta_0)]$ is a convex function of θ, for all θ_0. We can express this convexity condition using the equivalent first-order characterization:

$$\mathbb{E}_{z\sim\mathcal{D}(\theta)}[\ell(z;\theta_0)\nabla_\theta \log(p_\theta(z))]^\top(\theta' - \theta) \leq \mathbb{E}_{z\sim\mathcal{D}(\theta')}[\ell(z;\theta_0)] - \mathbb{E}_{z\sim\mathcal{D}(\theta)}[\ell(z;\theta_0)].$$

Since the mixture dominance condition holds for all θ, θ' and θ_0, we can set θ_0 equal to θ in the inequality above to conclude that

$$\mathbb{E}_{z\sim\mathcal{D}(\theta)}[\ell(z;\theta)\nabla_\theta \log(p_\theta(z))]^\top(\theta' - \theta) \leq \mathbb{E}_{z\sim\mathcal{D}(\theta')}[\ell(z;\theta)] - \mathbb{E}_{z\sim\mathcal{D}(\theta)}[\ell(z;\theta)].$$

Going back to equation (4.5), we see that a sufficient condition for $(\gamma - 2\epsilon\beta)$-convexity of the performative risk is

$$\mathbb{E}_{z\sim\mathcal{D}(\theta)}[\nabla_\theta \ell(z;\theta)]^\top(\theta' - \theta) + \frac{\gamma - 2\epsilon\beta}{2}\|\theta - \theta'\|_2^2 \leq \mathbb{E}_{z\sim\mathcal{D}(\theta')}\ell(z;\theta') - \mathbb{E}_{z\sim\mathcal{D}(\theta')}\ell(z;\theta).$$

By the assumption that the loss is γ-strongly convex in θ, we know

$$\mathbb{E}_{z\sim\mathcal{D}(\theta')}\ell(z;\theta') - \mathbb{E}_{z\sim\mathcal{D}(\theta')}\ell(z;\theta) \geq \mathbb{E}_{z\sim\mathcal{D}(\theta')}[\nabla_\theta \ell(z;\theta)]^\top(\theta' - \theta) + \frac{\gamma}{2}\|\theta - \theta'\|_2^2,$$

and thus we have further simplified the sufficient condition to

$$\mathbb{E}_{z\sim\mathcal{D}(\theta)}[\nabla_\theta \ell(z;\theta)]^\top(\theta' - \theta) - \mathbb{E}_{z\sim\mathcal{D}(\theta')}[\nabla_\theta \ell(z;\theta)]^\top(\theta' - \theta) \leq \frac{2\epsilon\beta}{2}\|\theta - \theta'\|_2^2.$$

Since the loss is β-smooth in z, we have that $\nabla_\theta \ell(z;\theta)^\top(\theta' - \theta)$ is $\beta\|\theta - \theta'\|_2$-Lipschitz in z. Now, we can use the fact that the distribution map is ϵ-sensitive to upper bound the left-hand side by applying the Kantorovich-Rubinstein duality theorem:

$$\mathbb{E}_{z\sim\mathcal{D}(\theta)}[\nabla_\theta \ell(z;\theta)]^\top(\theta' - \theta) - \mathbb{E}_{z\sim\mathcal{D}(\theta')}[\nabla_\theta \ell(z;\theta)]^\top(\theta' - \theta) \leq \epsilon\beta\|\theta - \theta'\|_2^2. \tag{4.6}$$

Therefore, we can conclude that the performative risk is $(\gamma - 2\epsilon\beta)$-convex.

Proof of Theorem 4.1.3.

Following the steps of Theorem 4.1.1, we know that $\mathrm{PR}(\theta)$ is λ-convex if and only if

$$\mathbb{E}_{z\sim\mathcal{D}(\theta)}[\nabla_\theta \ell(z;\theta)]^\top(\theta'-\theta) + \mathbb{E}_{z\sim\mathcal{D}(\theta)}[\ell(z;\theta)\nabla_\theta \log(p_\theta(z))]^\top(\theta'-\theta) + \frac{\lambda}{2}\|\theta-\theta'\|_2^2 \leq \mathrm{PR}(\theta') - \mathrm{PR}(\theta),$$

for all $\theta, \theta' \in \Theta$.

We now state a technical lemma which rephrases the conclusion of Proposition 4.5.5 in an equivalent way, deferring its proof to the end of this section.

Lemma 4.5.6. *Suppose that*

$$\mathbb{E}_{z\sim\mathcal{D}(\alpha\theta+(1-\alpha)\theta')}[g(z)] \leq \mathbb{E}_{z\sim\alpha\mathcal{D}(\theta)+(1-\alpha)\mathcal{D}(\theta')}[g(z)] - \frac{\alpha(1-\alpha)\gamma_z}{2}\mathbb{E}\|\Sigma(\theta-\theta')z_{\text{base}}+\mu(\theta-\theta')\|_2^2.$$

Then,

$$\mathbb{E}_{z\sim\mathcal{D}(\theta')}[g(z)] \geq \mathbb{E}_{z\sim\mathcal{D}(\theta)}[g(z)] + (\nabla_\theta \mathbb{E}_{z\sim\mathcal{D}(\theta)}[g(z)])^\top(\theta'-\theta) + \frac{\gamma_z}{2}\mathbb{E}\|\Sigma(\theta-\theta')z_{\text{base}}+\mu(\theta-\theta')\|_2^2.$$

Therefore, by Proposition 4.5.5 and Lemma 4.5.6, we know

$$\mathbb{E}_{z\sim\mathcal{D}(\theta)}[\ell(z;\theta)\nabla_\theta \log(p_\theta(z))]^\top(\theta'-\theta) \leq \mathbb{E}_{z\sim\mathcal{D}(\theta')}[\ell(z;\theta)] - \mathbb{E}_{z\sim\mathcal{D}(\theta)}[\ell(z;\theta)] - \frac{\gamma_z}{2}\mathbb{E}\|\Sigma(\theta-\theta')z_{\text{base}} + \mu(\theta-\theta')\|_2^2,$$

where we take $g(z) = \ell(z;\theta)$.

Thus it suffices to show

$$\mathbb{E}_{z\sim\mathcal{D}(\theta)}[\nabla_\theta \ell(z;\theta)]^\top(\theta'-\theta) + \frac{\lambda}{2}\|\theta-\theta'\|_2^2 \leq \mathbb{E}_{z\sim\mathcal{D}(\theta')}\ell(z;\theta') - \mathbb{E}_{z\sim\mathcal{D}(\theta')}\ell(z;\theta) + \frac{\gamma_z}{2}\mathbb{E}\|\Sigma(\theta-\theta')z_{\text{base}} + \mu(\theta-\theta')\|_2^2.$$

By the assumption that the loss is γ-strongly convex, we know

$$\mathbb{E}_{z\sim\mathcal{D}(\theta')}\ell(z;\theta') - \mathbb{E}_{z\sim\mathcal{D}(\theta')}\ell(z;\theta) \geq \mathbb{E}_{z\sim\mathcal{D}(\theta')}[\nabla_\theta \ell(z;\theta)]^\top(\theta'-\theta) + \frac{\gamma}{2}\|\theta-\theta'\|_2^2.$$

With this, we have simplified the sufficient condition for γ-convexity to

$$(\mathbb{E}_{z\sim\mathcal{D}(\theta)}[\nabla_\theta \ell(z;\theta)] - \mathbb{E}_{z\sim\mathcal{D}(\theta')}[\nabla_\theta \ell(z;\theta)])^\top(\theta'-\theta) \tag{4.7}$$

$$\leq \frac{\gamma-\lambda}{2}\|\theta-\theta'\|_2^2 + \frac{\gamma_z}{2}\mathbb{E}\|\Sigma(\theta-\theta')z_{\text{base}} + \mu(\theta-\theta')\|_2^2. \tag{4.8}$$

We bound the left-hand side by applying smoothness of the loss together with the Kantorovich-Rubinstein duality theorem; for this, we need a bound on $W(\mathcal{D}(\theta), \mathcal{D}(\theta'))$. We will use the bound implied by ϵ-sensitivity, as well as the bound implied by the following lemma.

Lemma 4.5.7. *Suppose that the distribution map $\mathcal{D}(\theta)$ forms a location-scale family* (4.3)*. Then,*

$$W(\mathcal{D}(\theta), \mathcal{D}(\theta')) \leq \mathbb{E}\|\Sigma(\theta - \theta')z_{\text{base}} + \mu(\theta - \theta')\|_2.$$

Proof of Lemma 4.5.7. By definition,

$$W(\mathcal{D}(\theta), \mathcal{D}(\theta')) = \inf_{\Pi(\mathcal{D}(\theta), \mathcal{D}(\theta'))} \mathbb{E}_{(z_\theta, z_{\theta'}) \sim \Pi(\mathcal{D}(\theta), \mathcal{D}(\theta'))}[\|z_\theta - z_{\theta'}\|_2],$$

where $\Pi(\mathcal{D}(\theta), \mathcal{D}(\theta'))$ denotes a coupling of $\mathcal{D}(\theta)$ and $\mathcal{D}(\theta')$. The simplest way to couple $\mathcal{D}(\theta)$ and $\mathcal{D}(\theta')$, or equivalently z_θ and $z_{\theta'}$, is to sample $z_{\text{base}} \sim \mathcal{D}$, and set $z_\theta = (\Sigma_0 + \Sigma(\theta))z_{\text{base}} + \mu_0 + \mu(\theta)$ and $z_{\theta'} = (\Sigma_0 + \Sigma(\theta'))z_{\text{base}} + \mu_0 + \mu(\theta')$. With this choice, $\|z_\theta - z_{\theta'}\|_2 = \|\Sigma(\theta - \theta')z_{\text{base}} + \mu(\theta - \theta')\|_2$, and hence

$$W(\mathcal{D}(\theta), \mathcal{D}(\theta')) \leq \mathbb{E}\|\Sigma(\theta - \theta')z_{\text{base}} + \mu(\theta - \theta')\|_2.$$

□

Therefore, the left-hand side in equation (4.7) can be bounded by

$$\mathbb{E}_{z \sim \mathcal{D}(\theta)}[\nabla_\theta \ell(z; \theta)]^\top(\theta' - \theta) - \mathbb{E}_{z \sim \mathcal{D}(\theta')}[\nabla_\theta \ell(z; \theta)]^\top(\theta' - \theta) \leq \beta\mathbb{E}\|\Sigma(\theta - \theta')z_{\text{base}} + \mu(\theta - \theta')\|_2\|\theta' - \theta\|_2,$$

but also by applying ϵ-sensitivity

$$\mathbb{E}_{z \sim \mathcal{D}(\theta)}[\nabla_\theta \ell(z; \theta)]^\top(\theta' - \theta) - \mathbb{E}_{z \sim \mathcal{D}(\theta')}[\nabla_\theta \ell(z; \theta)]^\top(\theta' - \theta) \leq \beta\epsilon\|\theta' - \theta\|_2^2.$$

Finally, to show $\lambda = \max\left\{\gamma - \beta^2/\gamma_z,\ \gamma + \gamma_z(\sigma_{\min}^2(\mu) + \sigma_{\min}^2(\Sigma)) - 2\beta\epsilon\right\}$-convexity it suffices to show both

$$\beta\mathbb{E}\|\Sigma(\theta - \theta')z_{\text{base}} + \mu(\theta - \theta')\|_2\|\theta' - \theta\|_2 \tag{4.9}$$

$$\leq \frac{\beta^2/\gamma_z}{2}\|\theta - \theta'\|_2^2 + \frac{\gamma_z}{2}\mathbb{E}\|\Sigma(\theta - \theta')z_{\text{base}} + \mu(\theta - \theta')\|_2^2 \tag{4.10}$$

and

$$\beta\epsilon\|\theta' - \theta\|_2^2 \leq \frac{2\beta\epsilon - \gamma_z(\sigma_{\min}^2(\mu) + \sigma_{\min}^2(\Sigma))}{2}\|\theta - \theta'\|_2^2 + \frac{\gamma_z}{2}\mathbb{E}\|\Sigma(\theta - \theta')z_{\text{base}} + \mu(\theta - \theta')\|_2^2. \tag{4.11}$$

By the AM-GM inequality, we have

$$\begin{aligned}&\beta\mathbb{E}\|\Sigma(\theta-\theta')z_{\text{base}}+\mu(\theta-\theta')\|_2\|\theta'-\theta\|_2\\&\leq\frac{1}{2}\frac{\beta^2}{\gamma_z}\|\theta'-\theta\|_2^2+\frac{\gamma_z}{2}\mathbb{E}\|\Sigma(\theta-\theta')z_{\text{base}}+\mu(\theta-\theta')\|_2^2,\end{aligned}$$

and so condition (4.9) follows.

For condition (4.11), we observe that

$$\begin{aligned}\mathbb{E}\|\Sigma(\theta-\theta')z_{\text{base}}+\mu(\theta-\theta')\|_2^2&=\mathbb{E}\|\Sigma(\theta-\theta')z_{\text{base}}\|_2^2+\|\mu(\theta-\theta')\|_2^2\\&=\text{Tr}\left(\Sigma(\theta-\theta')\Sigma_{z_{\text{base}}}\Sigma(\theta-\theta')^\top\right)+\|\mu(\theta-\theta')\|_2^2\\&=\|\Sigma_{z_{\text{base}}}^{1/2}\Sigma(\theta-\theta')^\top\|_F^2+\|\mu(\theta-\theta')\|_2^2.\end{aligned}$$

Applying $\sigma_{\min}(\Sigma)\|\theta-\theta'\|_2\leq\|\Sigma_{z_{\text{base}}}^{1/2}\Sigma(\theta-\theta')^\top\|_F$ and $\sigma_{\min}(\mu)\|\theta-\theta'\|_2\leq\|\mu(\theta-\theta')\|_2$ completes the proof of the theorem.

Proof of Lemma 4.5.6. The proof follows the standard argument for proving equivalent formulations of strong convexity.

First we show that $\mathbb{E}_{z\sim\mathcal{D}(\theta)}[g(z)]-\frac{\gamma_z}{2}\mathbb{E}\|\Sigma(\theta)z_{\text{base}}+\mu\theta\|_2^2$ is convex in θ. This follows because the difference

$$\mathbb{E}_{z\sim\mathcal{D}(\alpha\theta+(1-\alpha)\theta')}[g(z)]-\frac{\gamma_z}{2}\mathbb{E}\|\Sigma(\alpha\theta+(1-\alpha)\theta')z_{\text{base}}+\mu(\alpha\theta+(1-\alpha)\theta')\|_2^2$$

is upper bounded by:

$$\begin{aligned}&\leq\mathbb{E}_{z\sim\alpha\mathcal{D}(\theta)+(1-\alpha)\mathcal{D}(\theta')}[g(z)]-\frac{\alpha(1-\alpha)\gamma_z}{2}\mathbb{E}\|\Sigma(\theta-\theta')z_{\text{base}}+\mu(\theta-\theta')\|_2^2\\&-\frac{\gamma_z}{2}\mathbb{E}\|\Sigma(\alpha\theta+(1-\alpha)\theta')z_{\text{base}}+\mu(\alpha\theta+(1-\alpha)\theta')\|_2^2\\&=\mathbb{E}_{z\sim\alpha\mathcal{D}(\theta)+(1-\alpha)\mathcal{D}(\theta')}[g(z)]-\frac{\gamma_z}{2}\alpha^2\mathbb{E}\|\Sigma(\theta)z_{\text{base}}+\mu\theta\|_2^2\\&-\frac{\gamma_z}{2}(1-\alpha)^2\mathbb{E}\|\Sigma(\theta')z_{\text{base}}+\mu\theta'\|_2^2\\&+\frac{\gamma_z}{2}2\alpha(1-\alpha)\mathbb{E}(\Sigma(\theta)+\mu\theta)^\top(\Sigma(\theta')+\mu\theta')-\frac{\alpha(1-\alpha)\gamma_z}{2}\mathbb{E}\|\Sigma(\theta-\theta')z_{\text{base}}+\mu(\theta-\theta')\|_2^2\\&=\mathbb{E}_{z\sim\alpha\mathcal{D}(\theta)+(1-\alpha)\mathcal{D}(\theta')}[g(z)]-\frac{\gamma_z}{2}\alpha\mathbb{E}\|\Sigma(\theta)z_{\text{base}}+\mu\theta\|_2^2-\frac{\gamma_z}{2}(1-\alpha)\mathbb{E}\|\Sigma(\theta')z_{\text{base}}+\mu\theta'\|_2^2\\&=\alpha\left(\mathbb{E}_{z\sim\mathcal{D}(\theta)}[g(z)]-\frac{\gamma_z}{2}\mathbb{E}\|\Sigma(\theta)z_{\text{base}}+\mu\theta\|_2^2\right)\\&-(1-\alpha)\left(\mathbb{E}_{z\sim\mathcal{D}(\theta')}[g(z)]\frac{\gamma_z}{2}\mathbb{E}\|\Sigma(\theta')z_{\text{base}}+\mu\theta'\|_2^2\right).\end{aligned}$$

By the equivalent first-order characterization, this means that $\mathbb{E}_{z\sim\mathcal{D}(\theta')}[g(z)]$ is:

$$\begin{aligned}
&\geq \frac{\gamma_z}{2}\mathbb{E}\|\Sigma(\theta')z_{\text{base}} + \mu\theta'\|_2^2 + \mathbb{E}_{z\sim\mathcal{D}(\theta)}[g(z)] - \frac{\gamma_z}{2}\mathbb{E}\|\Sigma(\theta)z_{\text{base}} + \mu\theta\|_2^2 \\
&+ (\nabla_\theta \mathbb{E}_{z\sim\mathcal{D}(\theta)}[g(z)])^\top(\theta' - \theta) - \frac{\gamma_z}{2}2\mathbb{E}(\Sigma(\theta)z_{\text{base}} + \mu\theta)^\top(\nabla_\theta(\Sigma(\theta)z_{\text{base}} + \mu\theta))^\top(\theta' - \theta) \\
&\geq \frac{\gamma_z}{2}\mathbb{E}\|\Sigma(\theta')z_{\text{base}} + \mu\theta'\|_2^2 + \mathbb{E}_{z\sim\mathcal{D}(\theta)}[g(z)] - \frac{\gamma_z}{2}\mathbb{E}\|\Sigma(\theta)z_{\text{base}} + \mu\theta\|_2^2 \\
&+ (\nabla_\theta \mathbb{E}_{z\sim\mathcal{D}(\theta)}[g(z)])^\top(\theta' - \theta) - \gamma_z\mathbb{E}(\Sigma(\theta)z_{\text{base}} + \mu\theta)^\top(\Sigma(\theta' - \theta)z_{\text{base}} + \mu(\theta' - \theta)) \\
&= \mathbb{E}_{z\sim\mathcal{D}(\theta)}[g(z)] + (\nabla_\theta \mathbb{E}_{z\sim\mathcal{D}(\theta)}[g(z)])^\top(\theta' - \theta) + \frac{\gamma_z}{2}\mathbb{E}\|\Sigma(\theta - \theta')z_{\text{base}} + \mu(\theta - \theta')\|_2^2.
\end{aligned}$$

□

Remark 4.5.8. We note that the sensitivity parameter ϵ can be bounded in terms of the location and scale parameters for location-scale families. In particular, in showing condition (4.11), we saw that

$$\mathbb{E}\|\Sigma(\theta - \theta')z_{\text{base}} + \mu(\theta - \theta')\|_2^2 = \|\Sigma_{z_{\text{base}}}^{1/2}\Sigma(\theta - \theta')^\top\|_F^2 + \|\mu(\theta - \theta')\|_2^2.$$

If we then denote

$$\sigma_{\max}(\mu) = \max_{\|\theta\|_2=1} \|\mu\theta\|_2, \quad \sigma_{\max}(\Sigma) = \max_{\|\theta\|_2=1} \|\Sigma_{z_{\text{base}}}^{1/2}\Sigma(\theta)^\top\|_F,$$

we can see that $\mathbb{E}\|\Sigma(\theta - \theta')z_{\text{base}} + \mu(\theta - \theta')\|_2^2 \leq \sigma_{\max}^2(\mu)\|\theta - \theta'\|_2^2 + \sigma_{\max}^2(\Sigma)\|\theta - \theta'\|_2^2$. Combining this result with Lemma 4.5.7 and Jensen's inequality, we get that

$$W(\mathcal{D}(\theta), \mathcal{D}(\theta')) \leq \sqrt{\sigma_{\max}^2(\mu) + \sigma_{\max}^2(\Sigma)}\|\theta - \theta'\|_2,$$

and so $\epsilon \leq \sqrt{\sigma_{\max}^2(\mu) + \sigma_{\max}^2(\Sigma)}$.

Distinguishing between Static and Performative Variables

In many natural examples, the performative effects are only present in a subset of the variables that make up z. For example, in strategic classification, the performative effects are often only present in the strategically manipulated features, and not in the label.

For simplicity of exposition, we suppress this distinction between *performative* and *static* variables, that is, those whose distribution does not change for different $\mathcal{D}(\theta)$. However, the reader should think of all assumptions on z, such as strong convexity or various Lipschitz assumptions, as only having to apply to the performative variables, while the static ones can be averaged out. To give one example, suppose that $z = (z_s, z_p)$, where z_s

denotes the static variables and z_p denotes the performative ones. Using this distinction, the step in equation (4.6) would proceed as follows:

$$\begin{aligned}
&\mathbb{E}_{(z_s,z_p)\sim\mathcal{D}(\theta)}[\nabla_\theta \ell((z_s,z_p);\theta)]^\top(\theta'-\theta) - \mathbb{E}_{(z_s,z_p')\sim\mathcal{D}(\theta')}[\nabla_\theta \ell((z_s,z_p');\theta)]^\top(\theta'-\theta)\\
&= \mathbb{E}_{z_s}\left[\left(\mathbb{E}[\nabla_\theta \ell((z_s,z_p);\theta)|z_s]] - \mathbb{E}[\nabla_\theta \ell((z_s,z_p');\theta)|z_s]\right)^\top(\theta'-\theta)\right]\\
&\leq \mathbb{E}_{z_s}[\beta(z_s)\epsilon(z_s)]\,\|\theta-\theta'\|_2^2.
\end{aligned}$$

Here, $\beta(z_s)$ is the Lipschitz constant of $\nabla_\theta \ell((z_s,\cdot);\theta)$, and $\epsilon(z_s)$ is the sensitivity parameter of the distribution of z_p, conditional on z_s. As clear from the above example, stating all conditions and proofs while emphasizing this distinction is fairly cumbersome, so we opted for a simplified presentation. Similar calculations can be carried out for the rest of the proofs of the structural results.

Deferred Details of Two-Stage Algorithm for Location Families

We carefully review the problem setup and introduce the remaining assumptions. The distribution map $\mathcal{D}$ parameterizes a location family

$$z_\theta \sim \mathcal{D}(\theta) \;\Leftrightarrow\; z_\theta \overset{d}{=} z_{\text{base}} + \mu\theta,$$

where $z_{\text{base}} \sim \mathcal{D}_0$. We assume the base distribution $\mathcal{D}_0$ is zero-mean and subgaussian with parameter K. The loss function $\ell(z;\theta)$ is L_z-Lipschitz in z, L-Lipschitz and in θ, and β-smooth in (z,θ) in the sense that $\nabla\ell(z;\theta) \in \mathbb{R}^{m+d}$ is Lipschitz in (z,θ).

We also assume that $\lambda = \max\{\gamma - \beta^2/\gamma_z, \gamma - 2\epsilon\beta + \gamma_z\sigma_{\min}^2(\mu)\} > 0$, where γ and γ_z are the strong convexity parameters of the loss in θ and z, respectively. By Theorem 4.1.3, this implies that the performative risk is λ-strongly convex.

We assume that the performative optimum θ_{PO} is contained in a ball of radius R, so in the second stage we can set the domain of optimization to be $\Theta = \{\theta : \|\theta\|_2 \leq R\}$. Finally, we assume that the minimizer of the perturbed performative risk at the population level, $\hat{\theta} \in \arg\min_{\theta\in\Theta} \widehat{\text{PR}}(\theta)$ is contained in the interior of Θ with probability 1.

Theorem 4.5.9. *Under the preceding assumptions, if $n \geq \Omega(d + m + \log(1/\delta))$, then, with probability $1-\delta$, Algorithm 4.1 returns a point $\hat{\theta}_n$ such that*

$$\text{PR}(\hat{\theta}_n) - \text{PR}(\theta_{\text{PO}}) \leq O\left(\frac{d+m+\log(1/\delta)}{n} + \frac{1}{\delta n}\right).$$

Before proceeding to the proof of this result, we first state four auxiliary lemmas, which constitute the bulk of our analysis. The proofs of the lemmas are included in Section 4.5. The first lemma is a standard result about ordinary least-squares estimation.

Lemma 4.5.10. *If $n \geq \Omega(d + m + \log(1/\delta))$, then with probability $1 - \delta$,*

$$\|\mu - \hat{\mu}\| \leq O(\sqrt{\frac{(d+m) + \log(1/\delta)}{n}}).$$

The next lemma is a simple adaptation from Theorem 2 in [77] controlling the generalization gap of the empirical risk minimizer for strongly convex losses.

Lemma 4.5.11. *Suppose $\widehat{\mathrm{PR}}_n$ is $\hat{\lambda}$-strongly convex. Then, with probability at least $1 - \delta$,*

$$\widehat{\mathrm{PR}}(\hat{\theta}_n) - \widehat{\mathrm{PR}}(\hat{\theta}) \leq \frac{4(L_z\|\hat{\mu}\| + L)^2}{\delta\hat{\lambda}n}.$$

The next lemma controls the difference in gradients between the true performative risk PR and the perturbed performative risk $\widehat{\mathrm{PR}}$.

Lemma 4.5.12. *For any $\theta \in \Theta$,*

$$\|\nabla\mathrm{PR}(\theta) - \nabla\widehat{\mathrm{PR}}(\theta)\|_2^2 \leq O(\|\mu\|^2\|\mu - \hat{\mu}\|^2).$$

Finally, the last lemma shows that the smoothness assumptions on the loss ensure smoothness of the performative risk. Here, by β_θ-smoothness we mean that $\nabla_\theta\mathrm{PR}(\theta)$ is β_θ-Lipschitz.

Lemma 4.5.13. *Under the proceeding assumptions, the performative risk $\mathrm{PR}(\theta)$ is $\beta_\theta = O(\|\mu\|^2)$-smooth.*

With these lemmas in hand, we are now ready to prove Theorem 4.5.9.

Proof of Theorem 4.5.9. By assumption, the performative risk $\mathrm{PR}(\theta)$ is λ-strongly convex, for some $\lambda > 0$. This implies

$$\mathrm{PR}(\hat{\theta}_n) - \mathrm{PR}(\theta_{\mathrm{PO}}) \leq \frac{1}{2\lambda}\|\nabla\mathrm{PR}(\hat{\theta}_n)\|_2^2.$$

Since $\hat{\theta}_{\mathrm{PO}}$ is an interior minimizer of $\widehat{\mathrm{PR}}$, we know $\nabla\widehat{\mathrm{PR}}(\hat{\theta}_{\mathrm{PO}}) = 0$. Using $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$,

$$\begin{aligned}
\frac{1}{2\lambda}\|\nabla\mathrm{PR}(\hat{\theta}_n)\|_2^2 &= \frac{1}{2\lambda}\|\nabla\mathrm{PR}(\hat{\theta}_n) - \nabla\widehat{\mathrm{PR}}(\hat{\theta}_{\mathrm{PO}})\|_2^2 \\
&= \frac{1}{2\lambda}\|\nabla\mathrm{PR}(\hat{\theta}_n) - \nabla\widehat{\mathrm{PR}}(\hat{\theta}_n) + \nabla\widehat{\mathrm{PR}}(\hat{\theta}_n) - \nabla\widehat{\mathrm{PR}}(\hat{\theta}_{\mathrm{PO}})\|_2^2 \\
&\leq \frac{1}{\lambda}\|\nabla\mathrm{PR}(\hat{\theta}_n) - \nabla\widehat{\mathrm{PR}}(\hat{\theta}_n)\|_2^2 + \frac{1}{\lambda}\|\nabla\widehat{\mathrm{PR}}(\hat{\theta}_n) - \nabla\widehat{\mathrm{PR}}(\hat{\theta}_{\mathrm{PO}})\|_2^2. \qquad (4.12)
\end{aligned}$$

We bound each of these terms separately. For the first term, by Lemma 4.5.12,

$$\|\nabla \mathrm{PR}(\hat{\theta}_n) - \nabla \widehat{\mathrm{PR}}(\hat{\theta}_n)\|_2^2 \leq O(\|\mu\|^2 \|\mu - \hat{\mu}\|^2).$$

By Lemma 4.5.10, with probability $1 - \delta$, we can bound $\|\mu - \hat{\mu}\|^2 \leq O(\frac{d+m+\log(1/\delta)}{n})$, and thus

$$\|\nabla \mathrm{PR}(\hat{\theta}_n) - \nabla \widehat{\mathrm{PR}}(\hat{\theta}_n)\|_2^2 \leq O(\frac{d + m + \log(1/\delta)}{n}).$$

For the second term in equation (4.12), notice that $\lambda = \max\{\gamma - \beta^2/\gamma_z, \gamma - 2\epsilon\beta + \gamma_z \sigma_{\min}^2(\mu)\} > 0$ implies that $\widehat{\mathrm{PR}}$ is at least $\hat{\lambda} = \lambda - O(\frac{1}{\sqrt{n}})$-strongly convex. This follows because $|\sigma_{\min}(\mu) - \sigma_{\min}(\hat{\mu})| \leq \|\mu - \hat{\mu}\|$ by Weyl's inequality (see for example Theorem 3.3.16 in [71]), and $\widehat{\mathrm{PR}}$ is $O(\|\hat{\mu}\|)$-sensitive, so by Lemma 4.5.10, each term depending on ϵ or $\sigma_{\min}(\hat{\mu})$ is within $O(1/\sqrt{n})$ or $O(1/n)$ of the corresponding values for the non-perturbed risk PR.

Hence, when $n \geq \Omega(1/\lambda^2)$, the strong convexity parameter of the perturbed performative risk, $\hat{\lambda}$, is at least $\lambda/2$.

With this, we can apply the fact that $\hat{\theta}_{\mathrm{PO}}$ is an interior minimizer of $\widehat{\mathrm{PR}}$ by assumption to conclude that when $n \geq \Omega(1/\lambda^2)$,

$$\|\hat{\theta}_n - \hat{\theta}_{\mathrm{PO}}\|_2^2 \leq \frac{4}{\lambda}(\widehat{\mathrm{PR}}(\hat{\theta}_n) - \widehat{\mathrm{PR}}(\hat{\theta}_{\mathrm{PO}})).$$

Now, when $\widehat{\mathrm{PR}}$ is strongly convex, the finite-sample performative risk $\widehat{\mathrm{PR}}_n$ is also strongly convex because Theorem 4.1.3 does not depend on the base distribution $\mathcal{D}_0$, and $\widehat{\mathrm{PR}}_n$ is simply $\widehat{\mathrm{PR}}$ when the base distribution $\mathcal{D}_0$ is replaced with the uniform distribution on $\{z_1, \ldots, z_n\}$. Consequently, by Lemma 4.5.11, with probability $1 - \delta$,

$$\|\hat{\theta}_n - \hat{\theta}_{\mathrm{PO}}\|_2^2 \leq O(\widehat{\mathrm{PR}}(\hat{\theta}_n) - \widehat{\mathrm{PR}}(\hat{\theta}_{\mathrm{PO}})) \leq O(\frac{\|\hat{\mu}\|^2}{\delta n}).$$

By Lemma 4.5.13, $\widehat{\mathrm{PR}}$ is $O(\|\hat{\mu}\|^2)$-smooth. Applying the previous display then gives us,

$$\|\nabla \widehat{\mathrm{PR}}(\hat{\theta}_n) - \nabla \widehat{\mathrm{PR}}(\hat{\theta}_{\mathrm{PO}})\|_2^2 \leq O(\|\hat{\mu}\|^4 \|\hat{\theta}_n - \hat{\theta}_{\mathrm{PO}}\|_2^2) \leq O(\frac{\|\hat{\mu}\|^6}{\delta n}).$$

By the triangle inequality and repeated application of $(a + b)^2 \leq 2a^2 + 2b^2$, $\|\hat{\mu}\|^6 \leq 128\|\hat{\mu} - \mu\|^6 + 128\|\mu\|^6$. Therefore, the above term is $O(\|\mu\|^6/\delta n)$. Putting everything together with a union bound, we have shown that with probability $1 - \delta$, if $n \geq \Omega(d + m + \log(1/\delta))$, it holds that

$$\mathrm{PR}(\hat{\theta}_n) - \mathrm{PR}(\theta_{\mathrm{PO}}) \leq O(\frac{d + m + \log(1/\delta)}{n} + \frac{1}{\delta n}),$$

as desired. □

Proofs of Lemmas for Two-Stage Algorithm Analysis

The proof of Lemma 4.5.10 is essentially standard (see, e.g., [55]), but we include it for completeness.

Proof of Lemma 4.5.10. Define $Z \in \mathbb{R}^{n\times m}$ with rows z_i and $\Theta \in \mathbb{R}^{n\times d}$ with rows $\theta_i, 1 \le i \le n$. Then, $Z = \Theta\mu^\top + Z_0$, where $Z_0 \in \mathbb{R}^{n\times m}$ is a matrix with base samples from $\mathcal{D}_0$ as rows. Temporarily assume that $\Theta^\top\Theta$ is invertible; we will later condition on this event. Separately optimizing over each row of μ, we can write the least-squares estimator as

$$\hat{\mu}^\top = (\Theta^\top\Theta)^{-1}\Theta^\top Z.$$

Consequently, we can bound the estimation error as

$$\begin{aligned}\|\mu - \hat{\mu}\| = \|\mu^\top - \hat{\mu}^\top\| &= \|\mu^\top - (\Theta^\top\Theta)^{-1}\Theta^\top(\Theta\mu^\top + Z_0)\| \\ &= \|(\Theta^\top\Theta)^{-1}\Theta^\top Z_0\| \\ &\le \frac{1}{\lambda_{\min}(\Theta^\top\Theta)}\|\Theta^\top Z_0\|.\end{aligned}$$

Since $\theta_i \sim \mathcal{N}(0, I)$, $\Theta \in \mathbb{R}^{n\times d}$ has i.i.d. $\mathcal{N}(0,1)$ entries, and so $\Theta^\top\Theta$ is a standard Wishart matrix. The standard bound on the minimum eigenvalue of a Wishart matrix (see Theorem 4.6.1 in [84]) gives, with probability $1-\delta$,

$$\sqrt{\lambda_{\min}(\Theta^\top\Theta)} \ge \Omega(\sqrt{n} - \sqrt{d} - \sqrt{\log(1/\delta)}).$$

Therefore, if $n \ge \Omega(d + \log(2/\delta))$, then, with probability $1 - \delta/2$,

$$\sqrt{\lambda_{\min}(\Theta^\top\Theta)} \ge \Omega(\sqrt{n}/2). \tag{4.13}$$

Control of the second term, $\|\Theta^\top Z_0\|$, also follows from a standard covering argument followed by the Bernstein bound. Write $\Theta^\top Z_0 = \sum_{i=1}^n \theta_i (z_{\text{base}})_i^\top$. Let $\mathcal{B}^d$ and $\mathcal{B}^m$ denote the unit balls in $\mathbb{R}^d$ and $\mathbb{R}^m$, respectively. Then,

$$\|\Theta^\top Z_0\| = \sup_{x\in\mathcal{B}^d, y\in\mathcal{B}^m} x^\top(\sum_{i=1}^n \theta_i (z_{\text{base}})_i^\top)y = \sup_{x\in\mathcal{B}^d, y\in\mathcal{B}^m} \sum_{i=1}^n (x^\top\theta_i)((z_{\text{base}})_i^\top y).$$

Let $\mathcal{N}_\epsilon$, and $\mathcal{M}_\epsilon$ denote ϵ-coverings of $\mathcal{B}^d$ and $\mathcal{B}^m$, respectively. A volumetric bound gives $|\mathcal{N}_\epsilon| \le (1 + \frac{2}{\epsilon})^d$ and similarly $|\mathcal{M}_\epsilon| \le (1 + \frac{2}{\epsilon})^m$ (see Corollary 4.2.13 in [84]). Taking $\epsilon = 1/4$, $|\mathcal{N}_\epsilon| \le 9^d$ and $|\mathcal{M}_\epsilon| \le 9^m$. Approximating the supremum over the ϵ-nets gives

$$\|\Theta^\top Z_0\| \le 2 \max_{x\in\mathcal{N}_\epsilon, y\in\mathcal{M}_\epsilon} \sum_{i=1}^n (x^\top\theta_i)((z_0)_i^\top y).$$

Fix $x, y \in \mathcal{N}_\epsilon, \mathcal{M}_\epsilon$. Since $\theta_i \sim \mathcal{N}(0, I)$ and $\|x\|_2 = 1$, $x^\top \theta_i \sim \mathcal{N}(0, 1)$, which has subgaussian norm 1. Similarly, since $(z_{\text{base}})_i$ is subgaussian with parameter K and $\|y\|_2 = 1$, the marginal $(z_{\text{base}})_i^\top y$ is subgaussian with parameter K. Since z_{base} and θ are independent and zero-mean, the product $(x^\top \theta_i)((z_{\text{base}})_i^\top y)$ is zero-mean and subexponential with parameter K. Since each term is subexponential, by the Bernstein bound (see Theorem 2.8.1 in [84]), for any $t > 0$,

$$\mathbb{P}\left\{\sum_{i=1}^{n}(x^\top \theta_i)((z_{\text{base}})_i^\top y) > t/2\right\} \leq \exp(-c \min\left\{\frac{t^2}{nK^2}, \frac{t}{K}\right\}),$$

for some universal constant c. Taking a union bound over the ϵ-nets,

$$\mathbb{P}\left\{\|\Theta^\top Z_0\| > t\right\} \leq 9^{d+m} \exp(-c \min\left\{\frac{t^2}{nK^2}, \frac{t}{K}\right\}).$$

If $n \geq \Omega(d + m + \log(2/\delta))$, then with probability at least $1 - \delta/2$,

$$\|\Theta^\top Z_0\| \leq O(\sqrt{n((d+m) + \log(1/\delta))}). \tag{4.14}$$

Combining equations (4.13) and (4.14) with a union bound, if $n \geq \Omega(d + m + \log(1/\delta))$, then

$$\|\mu - \hat{\mu}\| \leq O\left(\sqrt{\frac{(d+m) + \log(1/\delta)}{n}}\right).$$

□

Proof of Lemma 4.5.12. Under the location-family parameterization, we can write

$$\text{PR}(\theta) = \mathbb{E}_{z \sim \mathcal{D}(\theta)} \ell(z; \theta) = \mathbb{E}_{z_0 \sim \mathcal{D}_0} \ell(z_0 + \mu\theta; \theta),$$

so the gradients are given by

$$\nabla \text{PR}(\theta) = \mathbb{E}_{z_0 \sim \mathcal{D}_0} \nabla \ell(z_0 + \mu\theta; \theta) \quad \text{and} \quad \nabla \widehat{\text{PR}}(\theta) = \mathbb{E}_{z_0 \sim \mathcal{D}_0} \nabla \ell(z_0 + \hat{\mu}\theta; \theta).$$

This representation allows us to write

$$\left\|\nabla \text{PR}(\theta) - \nabla \widehat{\text{PR}}(\theta)\right\|_2^2 = \left\|\begin{bmatrix} \mathbb{E}_{z_0 \sim \mathcal{D}_0} \\ \nabla \ell(z_0 + \mu\theta; \theta) - \nabla \ell(z_0 + \hat{\mu}\theta; \theta) \end{bmatrix}\right\|_2^2.$$

Applying the chain rule, together with the triangle-inequality, gives

$$\left\|\nabla \mathrm{PR}(\theta) - \nabla \widehat{\mathrm{PR}}(\theta)\right\|_2 \leq \left\|\begin{bmatrix} \mathbb{E}_{z_0 \sim \mathcal{D}_0} \\ \nabla_\theta \ell(z_0 + \mu\theta; \theta) - \nabla_\theta \ell(z_0 + \hat{\mu}\theta; \theta) \end{bmatrix}\right\|_2 + \left\|\begin{bmatrix} \mathbb{E}_{z_0 \sim \mathcal{D}_0} \\ \mu^\top \nabla_z \ell(z_0 + \mu\theta; \theta) - \hat{\mu}^\top \nabla_z \ell(z_0 + \hat{\mu}\theta; \theta) \end{bmatrix}\right\|_2.$$

We bound each of these terms separately. For the first term, β-smoothness in z immediately gives

$$\left\|\begin{bmatrix} \mathbb{E}_{z_0 \sim \mathcal{D}_0} \\ \nabla_\theta \ell(z_0 + \mu\theta; \theta) - \nabla_\theta \ell(z_0 + \hat{\mu}\theta; \theta) \end{bmatrix}\right\|_2 \leq \beta \|\mu\theta - \hat{\mu}\theta\|_2 \leq \beta \|\mu - \hat{\mu}\| \|\theta\|_2.$$

For the second term, adding and subtracting $\mu^\top \nabla_z \ell(z_0 + \hat{\mu}\theta; \theta)$ and then using the triangle inequality,

$$\begin{aligned} &\left\|\begin{bmatrix} \mathbb{E}_{z_0 \sim \mathcal{D}_0} \\ \mu^\top \nabla_z \ell(z_0 + \mu\theta); \theta) - \hat{\mu}^\top \nabla_z \ell(z_0 + \hat{\mu}\theta; \theta) \end{bmatrix}\right\|_2 \\ &\leq \|\mu\| \|\mathbb{E}_{z_0 \sim \mathcal{D}_0}[\nabla_z \ell(z_0 + \mu\theta); \theta) - \nabla_z \ell(z_0 + \hat{\mu}\theta; \theta)]\|_2 + \|\mu - \hat{\mu}\| \|\mathbb{E}_{z_0 \sim \mathcal{D}_0}[\nabla_z \ell(z_0 + \hat{\mu}\theta; \theta)]\|_2 \\ &\leq \beta \|\mu\| \|\mu - \hat{\mu}\| \|\theta\|_2 + L_z \|\mu - \hat{\mu}\|, \end{aligned}$$

where the last line used β-smoothness in z. Combining both pieces, we have

$$\left\|\nabla \mathrm{PR}(\theta) - \nabla \widehat{\mathrm{PR}}(\theta)\right\|_2 \leq ((\beta + \beta\|\mu\|)\|\theta\|_2 + L_z)\|\mu - \hat{\mu}\|.$$

Using the trivial bound $\|\theta\|_2 \leq R$, and then squaring both sides,

$$\|\nabla \mathrm{PR}(\hat{\theta}) - \nabla \widehat{\mathrm{PR}}(\hat{\theta})\|_2^2 \leq ((1 + \|\mu\|)\beta R + L_z)^2 \|\mu - \hat{\mu}\|^2.$$

□

Proof of Lemma 4.5.13. By applying the location family parameterization as in the proof of Lemma 4.5.12, we get

$$\|\nabla \mathrm{PR}(\theta) - \nabla \mathrm{PR}(\theta')\|_2 = \|\mathbb{E}_{z_0 \sim \mathcal{D}_0}[\nabla \ell(z_0 + \mu\theta; \theta) - \nabla \ell(z_0 + \mu\theta'; \theta')]\|_2.$$

Using the chain rule and the triangle inequality,

$$\begin{aligned} \|\nabla \mathrm{PR}(\theta) - \nabla \mathrm{PR}(\theta')\|_2 \leq\ & \|\mathbb{E}_{z_0 \sim \mathcal{D}_0} \nabla_\theta \ell(z_0 + \mu\theta; \theta) - \nabla_\theta \ell(z_0 + \mu\theta'; \theta')\|_2 \\ &+ \|\mathbb{E}_{z_0 \sim \mathcal{D}_0} \mu^\top \nabla_z \ell(z_0 + \mu\theta; \theta) - \mu^\top \nabla_z \ell(z_0 + \mu\theta'); \theta')\|_2. \end{aligned} \tag{4.15}$$

For the first term in equation (4.15), adding and subtracting $\nabla_\theta \ell(z + \mu\theta'; \theta)$ and using the triangle inequality gives $\|\mathbb{E}_{z_0 \sim \mathcal{D}_0}[\nabla_\theta \ell(z_0 + \mu\theta; \theta) - \nabla_\theta \ell(z_0 + \mu\theta'; \theta')]\|_2$

$$\begin{aligned}
&\leq \|\mathbb{E}_{z_0 \sim \mathcal{D}_0} \nabla_\theta \ell(z_0 + \mu\theta; \theta) - \nabla_\theta \ell(z_0 + \mu\theta'; \theta)\|_2 \\
&+ \|\mathbb{E}_{z_0 \sim \mathcal{D}_0} \nabla_\theta \ell(z_0 + \mu\theta'; \theta) - \nabla_\theta \ell(z_0 + \mu\theta'; \theta')\|_2 \\
&\leq \beta \|\mu\| \|\theta - \theta'\|_2 + \beta \|\theta - \theta'\|_2,
\end{aligned}$$

where we used Jensen's inequality and the assumption that $\nabla_\theta \ell(z; \theta)$ is β-Lipschitz in z (for the first term) and β-Lipschitz in θ (for the second term).

Now, for the second term in equation (4.15), similarly adding and subtracting $\mu^\top \nabla_z \ell(z + \mu\theta'; \theta)$ and using the triangle inequality gives

$$\begin{aligned}
&\|\mathbb{E}_{z_0 \sim \mathcal{D}_0}[\mu^\top \nabla_z \ell(z_0 + \mu\theta; \theta) - \mu^\top \nabla_z \ell(z_0 + \mu\theta'; \theta')]\|_2 \\
&\leq \|\mathbb{E}_{z_0 \sim \mathcal{D}_0} \mu^\top \nabla_z \ell(z_0 + \mu\theta; \theta) - \mu^\top \nabla_z \ell(z_0 + \mu\theta'; \theta)\|_2 \\
&+ \|\mathbb{E}_{z_0 \sim \mathcal{D}_0} \mu^\top \nabla_z \ell(z_0 + \mu\theta'; \theta) - \mu^\top \nabla_z \ell(z_0 + \mu\theta'; \theta')\|_2 \\
&\leq \beta \|\mu\|^2 \|\theta - \theta'\|_2 + \beta \|\mu\|^2 \|\theta - \theta'\|_2,
\end{aligned}$$

where we used $\nabla_z \ell(z; \theta)$ is β Lipschitz in z (for the first term) and β Lipschitz in θ (for the second term). This completes the proof. □

Experimental Details

Lastly, we elaborate on the experimental setup for the simulations presented in this chapter.

Data generation. We use the same strategic classification simulator as in the previous chapter. We consider two different values of the sensitivity parameter, $\epsilon \in \{0.0001, 100\}$, and set the magnitude of the regularizer to be $\lambda = 0.002$. We restrict the radius of the optimization domain to be 10, $\Theta = \{\theta : \|\theta\|_2 \leq 10\}$. This choice of parameters ensures that $\epsilon = 0.0001$ is below the critical threshold $\frac{\gamma}{2\beta}$, while $\epsilon = 100$ is above the threshold.

Algorithms. We compare the same four algorithms as the previous section.

1. **Two-stage procedure.** In the first stage, we deploy random $\theta_i \sim \mathcal{N}(0, I)$ and perform linear regression to estimate μ,

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \|z_i - \mu\theta_i\|^2$$

Then, having collected samples from the base distribution, we solve the proxy logistic regression objective offline by running gradient descent with a line search procedure until a tolerance criterion is met. In particular, we solve,

$$\arg\min_{\theta\in\Theta} \frac{1}{n} \sum_{j=n+1}^{2n} \ell(z_j + \hat{\mu}\theta; \theta),$$

where $\ell(z; \theta)$, is the regularized logistic regression objective, until the improvement between consecutive iterates is smaller than 1e-10.

2. **DFO.** We again run the derivative-free optimization procedure from Flaxman et al. [25]. We initialize $\theta_0 = \mathbf{0}$, use step-size sequence $1/t$, a batch size of 100 samples per-step, and set $\delta = 1$. We tried several other parameter configurations and found this one to perform best on this problem setting.

3. **Greedy SGD.** We run the greedy SGD variant with initial point $\theta_0 = \mathbf{0}$ and step-size sequence as suggested by our earlier theoretical analysis (Theorem 3.2.2).

4. **Lazy SGD.** We use the lazy SGD algorithm with initial point $\theta_0 = \mathbf{0}$ and k^2 collected samples in k-th update. As for greedy SGD, we use the step-size sequence suggested by the theory in Theorem 3.2.3.

Evaluation. We ran each algorithm for 50 trials, and in Figure 4.2, we compare the performative risk $\mathrm{PR}(\theta)$ of each algorithm as a function of the number of samples. For each sample size n, we bootstrap 95% confidence intervals over the 50 trials.

Chapter 5

In Search of Performative Optima II: Embracing the Multiplicity of Objectives

Throughout our presentation so far, we have assumed that the system designer is able to adequately encode the overarching goals of prediction into a single, fixed loss function ℓ that we then optimize via the performative risk. That is, we assume that this normative task of translating the subjective goals of prediction into a concrete, objective mathematical object has been resolved a priori. In particular, this translation happens before the learner has observed the data or begun to optimize their predictive model.

Often times, choosing the "right" objective may be a more challenging than finding the optimal predictive model for a specific loss function. As discussed previously (Chapter 2), a loss function is nothing but a way for people to express their preferences over (data, prediction) pairs. In supervised learning, we often choose loss functions such as the 01-loss (i.e., $\mathbf{1}\{\hat{y} \neq y\}$) or the squared loss (i.e., $(y - \hat{y})^2$) to indicate that we would like to find a predictive model that is *accurate*: it's predictions match future outcomes. This drive towards accuracy makes perfect sense if predictions are merely a tool to foreshadow future outcome and have no impact on the world.

In performative prediction, we are encouraged to think more broadly regarding the power of predictions. The performativity thesis of machine learning posits that any prediction that informs human decisions is an intervention that can actively change the data we observe. Consequently, a prediction is not just a way to *forecast* future outcomes accurately, but also to *steer* outcomes towards specific ends. In a medicine, risk predictions determine interventions and shape behavior. Therefore, we might want to find models that, for example, don't just tell us whether a person will experience a heart attack, but also minimize the likelihood that it occurs.

This distinction between *steering* and *forecasting* means that the process of translating

between the high-level goals of prediction and a concrete loss function can be quite an open ended process. The choice of loss in performative contexts is often inherently ambiguous and challenging.

In this chapter, we attempt to find technical solutions that directly embrace the multiplicity of objectives in performative prediction. In particular, drawing upon an exciting line of work in supervised learning [28, 29], we introduce the concept of a performative omnipredictor and illustrate how these solutions can be learned efficiently. Intuitively, a performative omnipredictor is a single predictive model that is simultaneously performatively optimal for many, diverse objectives. By diverse, we mean that these omnipredictors can be used to generate optimal predictions for qualitatively different, and possibly contradictory goals. In particular, they can induce performative optimal models for a forecasting loss (e.g, 01-loss) as well as for different steering losses (e.g., maximize the likelihood that the outcome is 1, as minimize the likelihood that it is 1).

In other words, performative omnipredictors are an "efficient menu" of optimal decision rules that enable the system-designer to "learn once, and choose the right objective later". They address this previous limitation of performative prediction regarding the choice of a single, fixed loss function and empower the decision-maker to flexibly decide on the high-level goals of prediction.

At first glance, this idea of a single model that is optimal for many diverse objectives seems like an impossibly strong goal. However, as we will now see, this concept can be provably achieved by extending an elegant analysis framework initially pioneered in the supervised learning context by [28]. The authors of this work establish a reduction between omniprediction and a notion of computational indistinguishability called Outcome Indistinguishability [18]. This reduction gives way to a simple algorithm for learning omnipredictors that requires only the simplest supervised learning primitives, such as the ability to solve weak learning problems. In this chapter, we draw upon these ideas to bring omniprediction into performative settings. And, in doing so, we establish a surprising reduction showing that omniprediction in performative prediction is in fact not much harder (computationally, or statistically speaking) than in seemingly simpler settings like supervised learning.

We prove all these results in the *outcome performativity* setting, where predictions only influence the distribution over outcomes y and the marginal distribution over features x is unaffected. The outcome performativity setup is a natural restriction of the performative prediction framework that matches the natural flow of time in prediction problems, as we will soon describe in more detail. However, this restriction means that are results regarding the possibilities of computing performative optima are generally incomparable to those found in previous chapters.

Remark 5.0.1 (a note regarding notation)**.** So far throughout this thesis, we have consider predictive models f_θ which are parametrized by a vector $\theta \in \mathbb{R}^d$. This was in large part motivated by the fact that we took a continuous optimization perspective on performative prediction whereby we considered the properties of the loss function with respect θ or the behavior of gradient based algorithms which directly operate in θ.

The results in this chapter are more "discrete" in flavor. We will consider predictive models, or classifiers, that have outputs in a discrete set $\hat{\mathcal{Y}}$ are not necessarily cleanly parametrized by a vector θ (e.g., decision-trees). In line with common practices in the computational learning theory literature, we will refer to predictive models as decision-rules or hypothesis functions h belonging to some class $\mathcal{H}$. Consequently, h will be the input to the distribution map $\mathcal{D}(\cdot)$ and we will use $\mathcal{D}(h)$ to denote the joint distribution over pairs (x, y) when predictions are made according to the decision rule h.

5.1 Outcome Performativity

To begin, we formally define a special case of the performative prediction setting, which we call *outcome performativity*. Outcome performativity focuses on the effects of local decisions on individuals' outcomes, rather than the effect of broader policy on the distribution of individuals.

For instance, early warning systems (which we will study in more detail in the second part of this thesis) is well captured by the outcome performativity. For a given a student, the EWS prediction they receive affects their future graduation outcome, but does not influence their demographic features or historical test scores. In other words, we narrow our attention to the performative effects of the prediction itself $h(x)$ on the conditional distribution over outcomes y, rather than the effects of the decision rule h on the distribution as a whole $\mathcal{D}(h)$. This specialization of performativity still captures many important decision-making problems, but gives us additional structure to sidestep some of the difficulties highlighted in previous chapters regarding the difficulty of finding performative optima.

On a technical level, outcome performativity imagines a data generating process over triples $(x, \hat{y}, y^*)$ where $x \sim \mathcal{D}$ is sampled from a *static* distribution over inputs, then a prediction or decision $\hat{y} \in \hat{\mathcal{Y}}$ is selected (possibly as a function of x), and finally the true outcome $y^* \in \mathcal{Y}$ is sampled conditioned on x and $\hat{y}$. Throughout this chapter, we focus on binary outcomes $\mathcal{Y} = \{0, 1\}$.[1] In this setting, the outcome performativity assumption

[1] In general, outcome performativity could be defined for larger outcome domains. Handling such domains is certainly possible, but technical. We restrict our attention to binary outcomes to focus on conceptual issues. Given this restriction, we use $y \sim p(x, \hat{y})$ as shorthand for $y \sim \text{Ber}[p(x, \hat{y})]$.

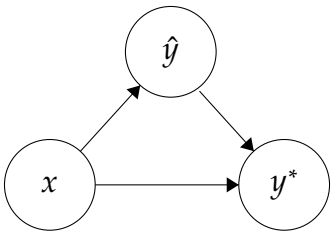


Figure 5.1: Causal graphical representation of the outcome performativity data generating process.

posits the existence of an underlying probability function,

$$p^* : \mathcal{X} \times \hat{\mathcal{Y}} \to [0, 1],$$

where for a given individual $x \in \mathcal{X}$ and decision $\hat{y} \in \hat{\mathcal{Y}}$, the true outcome y^* is sampled as a Bernoulli with parameter $p^*(x, \hat{y})$. We refer to the true outcome distribution p^* as *Nature*.

By asserting a fixed "ground truth" probability function, the outcome performativity framework does not allow for arbitrary distributional responses and limits the generality of the approach. For instance, outcome performativity does not capture strategic classification. But importantly, by refining the model of performativity, there is hope that we may sidestep the general hardness of learning optimal performative predictors.

5.2 Performative Omniprediction

We begin by observing that under outcome performativity, the true probability function p^* suggests an optimal decision rule $f_\ell^* : \mathcal{X} \to \hat{\mathcal{Y}}$ for any loss ℓ. In our setting, p^* governs the outcome distribution, so given an input $x \in \mathcal{X}$, the optimal decision $f_\ell^*(x)$ is determined by a simple, univariate optimization procedure over a discrete set $\hat{\mathcal{Y}}$:

$$f_\ell^*(x) \in \arg\min_{\hat{y} \in \hat{\mathcal{Y}}} \mathbb{E}_{y^* \sim p^*(x, \hat{y})}[\ell(x, \hat{y}, y^*)]. \tag{5.1}$$

Note that the decision rule $f_\ell^*(x)$ minimizes the loss pointwise for $x \in \mathcal{X}$. Consequently, averaging over any static, feature distribution $\mathcal{D}$, the decision rule f_ℓ^* is performative optimal for *any* hypothesis class $\mathcal{H}$, loss ℓ, and marginal distribution $\mathcal{D}$:

$$\mathbb{E}_{\substack{x \sim \mathcal{D} \\ y^* \sim p^*(x, f_\ell^*(x))}} [\ell(x, f_\ell^*(x), y^*)] \leq \min_{h \in \mathcal{H}} \mathbb{E}_{\substack{x \sim \mathcal{D} \\ y^* \sim p^*(x, h(x))}} [\ell(x, h(x), y^*)].^2$$

While the existence of p^* implies the existence of optimal decision rules under outcome performativity, we make no assumptions about the learnability of p^*. In general, the function p^* may be arbitrarily complex, so learning (or even representing!) p^* may be infeasible, both computationally and statistically.

Still, the above analysis reveals the power of modeling the probability function

$$p^* : \mathcal{X} \times \hat{\mathcal{Y}} \to [0, 1].$$

The optimal probability function p^* encodes the optimal decision rule f_ℓ^* *for every loss function* ℓ. This perspective raises a concrete technical question: short of learning p^*, can we learn a probability function $\tilde{p} : \mathcal{X} \times \hat{\mathcal{Y}} \to [0, 1]$ that suggests an optimal decision rule, via simple post-processing, for many different objectives?

Recent work of [29] studied the analogous question in the context of supervised learning (without performativity), formalizing a solution concept which they call *omniprediction*. Intuitively, an omnipredictor is a single probability function $\tilde{p}$ that suggests an optimal decision rule for many different loss functions $\mathcal{L}$. The work of [29] and follow-up work of [28] demonstrate—rather surprisingly—that omniprediction in supervised learning is broadly a feasible concept. For a variety of choices of loss classes $\mathcal{L}$ (e.g., Lipschitz losses or convex losses), it is possible to learn an efficient predictor $\tilde{p}$ that gives optimal decisions for any loss $\ell \in \mathcal{L}$.

In this chapter, we generalize omniprediction to the outcome performative setting. As a solution concept, *performative omniprediction* directly addresses the limiting assumption in performative prediction that the loss ℓ is known and fixed. Given a performative omnipredictor, a decision-maker can explore the consequences of optimizing for different losses, balancing the desire for forecasting and steering, as they see fit.

Technically, given a predictor $\tilde{p}$, we define $\tilde{f}_\ell : \mathcal{X} \to \hat{\mathcal{Y}}$ to be the optimal decision rule, that acts as if outcomes are governed by $\tilde{p}$.

$$\tilde{f}_\ell(x) \in \arg\min_{\hat{y} \in \hat{\mathcal{Y}}} \mathbb{E}_{\tilde{y} \sim \tilde{p}(x, \hat{y})}[\ell(x, \hat{y}, \tilde{y})]$$

We emphasize that, for any loss ℓ, the decision rule $\tilde{f}_\ell(x)$ is an efficient post-processing of the predictions given by $\tilde{p}(x, \hat{y})$ for $\hat{y} \in \hat{\mathcal{Y}}$. A performative omnipredictor is a model of nature $\tilde{p} : \mathcal{X} \times \hat{\mathcal{Y}} \to [0, 1]$ that induces a corresponding decision rule $\tilde{f}_\ell$ that is performatively optimal over a collection of losses $\ell \in \mathcal{L}$:

Definition 5.2.1 (Performative Omniprediction)**.** For input distribution $\mathcal{D}$, collection of loss functions $\mathcal{L}$, hypothesis class $\mathcal{H}$, and $\epsilon \geq 0$, a predictor $\tilde{p} : \mathcal{X} \times \hat{\mathcal{Y}} \to [0, 1]$ is an $(\mathcal{L}, \mathcal{H}, \epsilon)$-performative omnipredictor over $\mathcal{D}$ if for all $\ell \in \mathcal{L}$, the optimal decision rule $\tilde{f}_\ell$ is $(\ell, \mathcal{H}, \epsilon)$-performative optimal.

Omniprediction is a very strong solution concept. Whereas the optimal decision rule typically depends intimately on the chosen loss, an omnipredictor needs to encode the optimal decision rule for every loss in $\mathcal{L}$, even if these losses encode very different preferences over predictions. It is not hard to see that the optimal predictor (i.e, Nature's p^*) is an omnipredictor for any hypothesis and loss class.

Corollary 5.2.2. *For any input distribution $\mathcal{D}$, collection of loss functions $\mathcal{L}$, and hypothesis class $\mathcal{H}$, the optimal predictor $p^* : \mathcal{X} \times \hat{\mathcal{Y}} \to [0,1]$ is an $(\mathcal{L}, \mathcal{H}, 0)$-performative omnipredictor over $\mathcal{D}$.*

This corollary follows directly from the fact that the optimal predictor p^* gives the true probability law governing the performative outcome distribution. Still, as we discussed previously, the optimal predictor may be of arbitrary complexity and is generally inaccessible. The question remains whether *efficient* performative omnipredictors exist, and if so, how to learn them. To attack this question, we introduce a generalization of the outcome indistinguishability framework to the outcome performativity setting.

Performative Outcome Indistinguishability

Outcome Indistinguishability (OI) was introduced by [18] as an alternative paradigm for supervised learning. Rather than focusing on loss minimization, OI formalizes learning as a computational indistinguishability condition. In this view, a predictor should produce outcomes that are indistinguishable from Nature's outcome distribution. While OI can encode classic learning goals like loss minimization, the abstraction is quite generic and amenable to modern supervised learning desiderata, like fairness [34] and distributional robustness [43].

Here, we propose an indistinguishability definition for the performative world. This definition extends what [28] refer to as Hypothesis OI in the supervised setting. Here, we propose an indistinguishability definition for the performative world.[3] This definition extends what [28] refer to as Hypothesis OI in the supervised setting.

Definition 5.2.3 (Performative OI)**.** For input distribution $\mathcal{D}$, collection of losses $\mathcal{L}$, hypothesis class $\mathcal{H}$, and $\epsilon \geq 0$, a predictor $\tilde{p} : \mathcal{X} \times \hat{\mathcal{Y}} \to [0,1]$ is $(\mathcal{L}, \mathcal{H}, \epsilon)$-performative outcome indistinguishable (POI) over $\mathcal{D}$ if for all $\ell \in \mathcal{L}$ and all $h \in \mathcal{H}$,

$$\left| \mathbb{E}_{\substack{x \sim \mathcal{D} \\ y^* \sim p^*(x, h(x))}} [\ell(x, h(x), y^*)] - \mathbb{E}_{\substack{x \sim \mathcal{D} \\ \tilde{y} \sim \tilde{p}(x, h(x))}} [\ell(x, h(x), \tilde{y})] \right| \leq \epsilon$$

[3] In its original formulation, OI is a hierarchy of related notions. We generalize the framework to our setting, focusing on notions of performative OI that will imply performative omniprediction. Understanding a full generalization of the OI framework to the performative setting is an interesting question for future investigations.

In this definition, we fix our collection of distinguishers to be parameterized by a collection of loss functions and a hypothesis class. The POI condition states that, even when the outcome distribution can depend nontrivially on the hypothesis value $h(x)$, the outcomes y^* and $\tilde{y}$ are indistinguishable, as measured by the expected loss of each hypothesis. Note that the distinguishers take as input the individual x, the decision $h(x)$, and either Nature's outcome y^* or the modeled outcome $\tilde{y}$. In particular, these distinguishers do not receive access to the predictions $\tilde{p}(x, h(x))$ themselves.[4]

As a step towards obtaining omniprediction, we require indistinguishability between $\tilde{p}$ and p^*, not just under the reference decision rules h, but also under the optimal decision rules $\tilde{f}_\ell$ derived from $\tilde{p}$. This motivates the notion of Performative Decision OI, which extends the idea of decision calibration, introduced in [89], and decision OI, introduced in [28], to the performative setting.

Definition 5.2.4 (Performative Decision OI)**.** For input distribution $\mathcal{D}$, collection of loss functions $\mathcal{L}$, and $\epsilon \geq 0$, a predictor $\tilde{p} : \mathcal{X} \times \hat{\mathcal{Y}} \to [0, 1]$ is $(\mathcal{L}, \epsilon)$-performative decision outcome indistinguishable (DOI) over $\mathcal{D}$ if for all $\ell \in \mathcal{L}$,

$$\left| \mathbb{E}_{\substack{x \sim \mathcal{D} \\ y^* \sim p^*(x, \tilde{f}_\ell(x))}} [\ell(x, \tilde{f}_\ell(x), y^*)] - \mathbb{E}_{\substack{x \sim \mathcal{D} \\ \tilde{y} \sim \tilde{p}(x, \tilde{f}_\ell(x))}} [\ell(x, \tilde{f}_\ell(x), \tilde{y})] \right| \leq \epsilon.$$

Operationally, DOI allows us to sample outcomes $\tilde{y} \sim \tilde{p}(x, \tilde{f}_\ell(x))$ from our model of Nature, evaluate the expected loss of $\ell(x, \tilde{f}_\ell(x), \tilde{y})$, and be confident that it is close to the loss on outcomes sampled from Nature $y^* \sim p^*(x, \tilde{f}_\ell(x))$.

Note that, technically, the indistinguishability conditions in Performative OI and Performative Decision OI look the same, but just refer to different hypothesis classes; that is, $(\mathcal{L}, \epsilon)$-Performative Decision OI can be phrased as $(\mathcal{L}, \{\tilde{f}_\ell : \ell \in \mathcal{L}\}, \epsilon)$-Performative OI. We make a distinction between these notions because, semantically, the hypothesis class $\{\tilde{f}_\ell : \ell \in \mathcal{L}\}$ is derived from the predictor $\tilde{p}$, whereas $h \in \mathcal{H}$ is independent of $\tilde{p}$. As we will see later, this semantic difference manifests as a concrete difference in the computational complexity of achieving each notion of indistinguishability.

Performative Omniprediction via OI

With these definitions in place, we can prove our first main result: performative omniprediction from performative outcome indistinguishability. One of the main benefits of studying the problem from the indistinguishability lens is that it enables an especially

[4] In the language of [18], this notion corresponds to the "No-Access" level of the OI hierarchy. In principle, we could also extend the upper levels to the performative setting as well. We comment this issue further within our discussion of performative calibration in Section 5.5.

clean and simple analysis. The proof strategy we employ here follows the proof of omniprediction in the supervised learning world by [28]. Curiously, the proof only needs one direction of the indistinguishability inequalities.

Theorem 5.2.5. *Fix an input distribution $\mathcal{D}$, collection of losses $\mathcal{L}$, hypothesis class $\mathcal{H}$, and $\epsilon \geq 0$. Suppose that $\tilde{p} : \mathcal{X} \times \hat{\mathcal{Y}} \to [0,1]$ is $(\mathcal{L}, \epsilon)$-performative decision OI and $(\mathcal{L}, \mathcal{H}, \epsilon)$-performative OI. Then, $\tilde{p}$ is a $(\mathcal{L}, \mathcal{H}, 2\epsilon)$-performative omnipredictor.*

Proof. The proof exploits the fact that for each loss $\ell \in \mathcal{L}$, $\tilde{f}_\ell$ is the optimal decision rule for ℓ under $\tilde{p}$. Fix a loss $\ell \in \mathcal{L}$. First, we upper bound the loss achieved by $\tilde{f}_\ell$ on real outcomes y^* in terms of the loss on modeled outcomes $\tilde{y}$. Under $(\mathcal{L}, \epsilon)$-performative decision OI,

$$\mathbb{E}_{\substack{x\sim\mathcal{D} \\ y^*\sim p^*(x,\tilde{f}_\ell(x))}}[\ell(x, \tilde{f}_\ell(x), y^*)] \leq \mathbb{E}_{\substack{x\sim\mathcal{D} \\ \tilde{y}\sim\tilde{p}(x,\tilde{f}_\ell(x))}}[\ell(x, \tilde{f}_\ell(x), \tilde{y})] + \epsilon.$$

Next, we relate the expected loss achieved by $\tilde{f}_\ell$ on modeled outcomes $\tilde{y} \sim \tilde{p}(x, \tilde{f}_\ell(x))$ versus that of other decision rules h. By its definition, $\tilde{f}_\ell(x)$ is the optimal decision over any $\hat{y} \in \hat{\mathcal{Y}}$ for the loss $\ell(x, \hat{y}, \tilde{y})$ under $\tilde{y} \sim \tilde{p}(x, \hat{y})$. So, averaging over the distribution on inputs $x \sim \mathcal{D}$, the loss of $\tilde{f}_\ell$ is upper bounded by the loss of any other decision rule h, and in particular those in $\mathcal{H}$:

$$\mathbb{E}_{\substack{x\sim\mathcal{D} \\ \tilde{y}\sim\tilde{p}(x,\tilde{f}_\ell(x))}}[\ell(x, \tilde{f}_\ell(x), \tilde{y})] \leq \mathbb{E}_{\substack{x\sim\mathcal{D} \\ \tilde{y}\sim\tilde{p}(x,h(x))}}[\ell(x, h(x), \tilde{y})].$$

Finally, by $(\mathcal{L}, \mathcal{H}, \epsilon)$-POI, we upper bound the loss achieved by h on real outcomes y^* by that achieved on modeled outcomes $\tilde{y}$.

$$\mathbb{E}_{\substack{x\sim\mathcal{D} \\ \tilde{y}\sim\tilde{p}(x,h(x))}}[\ell(x, h(x), \tilde{y})] \leq \mathbb{E}_{\substack{x\sim\mathcal{D} \\ y^*\sim p^*(x,h(x))}}[\ell(x, h(x), y^*)] + \epsilon.$$

Combining these three inequalities,

$$\mathbb{E}_{\substack{x\sim\mathcal{D} \\ y^*\sim p^*(x,\tilde{f}_\ell(x))}}[\ell(x, \tilde{f}_\ell(x), y^*)] \leq \mathbb{E}_{\substack{x\sim\mathcal{D} \\ y^*\sim p^*(x,h(x))}}[\ell(x, h(x), y^*)] + 2\epsilon,$$

so $\tilde{p}$ is a $(\mathcal{L}, \mathcal{H}, 2\epsilon)$-performative omnipredictor. □

5.3 Universal Adaptability

In addition to minimizing expected risk, performative omniprediction can also be viewed as a guarantee of robustness. So far, we've seen how a performative omnipredictor induces optimal predictions $\hat{y}$ even if these predictions lead to endogenous shifts in the

distribution over outcomes y^*. In this section, we argue that with little additional work, the OI framework can be adapted to yield performative omnipredictors that are robust to exogenous shifts in the marginal distribution over individuals x.[5]

The results here build on the recent work of [43], who introduced a notion of *universal adaptability* in the context of statistical inference problems. In our context, universal adaptability may be interpreted as a guarantee that the performative omnipredictor properties hold, not only on the original input distribution $\mathcal{D}$, but also on a broad family of shifts of this input distribution $\mathcal{D}$. In particular, we show that by augmenting the class of loss functions, we can learn an outcome prediction model $\tilde{p}$ that can handle exogenous shifts in the input distribution, while still maintaining performative optimality.

We parameterize universal adaptability by a class of importance weight functions $\mathcal{W} \subseteq \{\mathcal{X} \to \mathbb{R}_{\geq 0}\}$. For a base input distribution $\mathcal{D}$, we define a corresponding collection of shifted distributions $\mathcal{D}_{\mathcal{W}}$ to be the set of distributions reachable after reweighting the probabilities in $\mathcal{D}$ by some $\omega \in \mathcal{W}$.

$$\mathcal{D}_{\mathcal{W}} = \{\mathcal{D}_\omega : \omega \in \mathcal{W},\ \mathrm{supp}(\mathcal{D}_\omega) \subseteq \mathrm{supp}(\mathcal{D})\},$$
$$\text{where } \forall x \in \mathrm{supp}(\mathcal{D}_\omega),\ \ \mathcal{D}_\omega(x) = \omega(x) \cdot \mathcal{D}(x)$$

Note that to yield a valid probability distribution $\mathcal{D}_\omega$ it is necessary and sufficient that the importance weight function ω have unit weight over $\mathcal{D}$; that is, for any $\omega \in \mathcal{W}$, $\mathbb{E}_{x\sim\mathcal{D}}[\omega(x)] = 1$.[6] Given an importance weight class $\mathcal{W}$, we say that an omnipredictor is universally adaptable if it is an omnipredictor over any $\mathcal{D}_\omega \in \mathcal{D}_{\mathcal{W}}$.

Definition 5.3.1 (Universal Adaptability)**.** For input distribution $\mathcal{D}$, weight class $\mathcal{W}$, collection of losses $\mathcal{L}$, hypothesis class $\mathcal{H}$, and $\epsilon \geq 0$, a performative omnipredictor is $\mathcal{W}$-universally adaptable over $\mathcal{D}$ if $\tilde{p}$ is an $(\mathcal{L}, \mathcal{H}, \epsilon)$-performative omnipredictor over every $\mathcal{D}_\omega \in \mathcal{D}_{\mathcal{W}}$.

Note that universal adaptability guarantees robustness under *exogeneous* shifts in the marginal distribution over $\mathcal{X}$, not under *endogeneous* shifts in the input distribution induced by the act of prediction. The distributional robustness is with respect to shifts that are defined in advance, independent of the chosen decision rule. Under the guarantees of universal adaptability, the prevalence of various individuals may vary, but the response of any specific individual x to a prediction $\hat{y}$, as measured by the distribution p^* governing

[5]By endogenous we mean that the distribution shift is caused by the act of prediction itself, which is considered in the outcome performativity framework. Exogenous shifts are not influence by predictions. They refer to changes in the data distribution caused by factors like a change in external environment, or the passage of time.

[6]Other properties of $\mathcal{W}$ will affect whether universal adaptability is feasible, but not its definition. We discuss these issues further in Section 5.4.

the outcome y^*, remains the same. Intuitively, this type of robustness is the best that we can hope for without explicitly modeling how the predictions $\hat{y} \in \hat{\mathcal{Y}}$ change the distribution over individuals $x \in \mathcal{X}$, which $\tilde{p}$ does not model. If predictions $\hat{y}$ affect both x and y^*, it is not at all obvious to us what invariant property of Nature we should choose to model. We believe these are important questions for future work.

One consequence of this adaptability definition is that any model $\tilde{p}$ that is an omnipredictor for a class of distributions $\mathcal{D}_{\mathcal{W}}$ must also be an omnipredictor for any mixture distribution with components drawn from this class. We say that a distribution $\mathcal{D}_m$ is a mixture distribution if for all $x \in \mathcal{X}$,

$$\Pr_{\mathcal{D}_m}[X = x] = \sum_{\omega} \lambda_\omega \Pr_{\mathcal{D}_\omega}[X = x] \text{ where } \mathcal{D}_\omega \in \mathcal{D}_{\mathcal{W}},\ \lambda_\omega \geq 0 \text{ for all } \omega \text{ and } \sum_{\omega} \lambda_\omega = 1.$$

We denote by $\mathsf{mixt}(\mathcal{D}_{\mathcal{W}})$ the set of all such mixture distributions $\mathcal{D}_m$.

Proposition 5.3.2. *Let $\mathcal{D}_{\mathcal{W}}$ be a set of distributions over $\mathcal{X}$. If $\tilde{p}$ is a $(\mathcal{L}, \mathcal{H}, \epsilon)$-performative omnipredictor over every $\mathcal{D}_\omega \in \mathcal{D}_{\mathcal{W}}$, then it is also a $(\mathcal{L}, \mathcal{H}, \epsilon)$-performative omnipredictor over every $\mathcal{D}_m$ in $\mathsf{mixt}(\mathcal{D}_{\mathcal{W}})$.*

Proof. Fix a loss $\ell \in \mathcal{L}$, a hypothesis $h \in \mathcal{H}$ and a distribution $\mathcal{D}_m$ in $\mathsf{mixt}(\mathcal{D}_{\mathcal{W}})$. Then,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}_m \\ y^*\sim p^*(x,\tilde{f}_\ell(x))}}[\ell(x, \tilde{f}_\ell(x), y^*)] &= \sum_{\omega} \lambda_\omega \cdot \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}_i \\ y^*\sim p^*(x,\tilde{f}_\ell(x))}}[\ell(x, \tilde{f}_\ell(x), y^*)] \\
&\leq \sum_{\omega} \lambda_\omega \cdot \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}_\omega \\ y^*\sim p^*(x,h(x))}}[\ell(x, h(x), y^*)] + \sum_{\omega} \lambda_\omega \epsilon \\
&= \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}_m \\ y^*\sim p^*(x,h(x))}}[\ell(x, h(x), y^*)] + \epsilon
\end{aligned}$$

The first line follows by expanding the definition of the mixture distribution and the second by the omniprediction guarantee on mixture components. In the last line we again applied the definition of a mixture and the fact that the λ_ω sum to 1. Because the inequalities hold for every $h \in \mathcal{H}$, it must be the case that $\tilde{p}$ is an $(\mathcal{L}, \mathcal{H}, \epsilon)$-omnipredictor for every mixture distribution . □

We establish universal adaptability for performative omnipredictors by augmenting the loss class $\mathcal{L}$ using the weight class $\mathcal{W}$. Specifically, we define the augmented loss class $\mathcal{L}_{\mathcal{W}}$ as the class of losses $\ell \in \mathcal{L}$ reweighted by importance weight functions $\omega \in \mathcal{W}$.

$$\mathcal{L}_{\mathcal{W}} = \{\ell_\omega : \ell \in \mathcal{L},\ \omega \in \mathcal{W}\}$$
$$\text{where } \forall x \in \mathcal{X}, \hat{y} \in \hat{\mathcal{Y}}, y \in \mathcal{Y}: \quad \ell_\omega(x, \hat{y}, y) = \omega(x) \cdot \ell(x, \hat{y}, y)$$

With this class of losses in place, we argue that universally-adaptable performative omniprediction is, again, a consequence of performative outcome indistinguishability.

Proposition 5.3.3. *For a base input distribution $\mathcal{D}$, weight class $\mathcal{W}$, collection of losses $\mathcal{L}$, hypothesis class $\mathcal{H}$, and $\epsilon \geq 0$, if a predictor $\tilde{p}$ is $(\mathcal{L}_{\mathcal{W}}, \mathcal{H}, \epsilon)$-performative OI and $(\mathcal{L}_{\mathcal{W}}, \epsilon)$-performative decision OI over $\mathcal{D}$, then $\tilde{p}$ is an $(\mathcal{L}, \mathcal{H}, 2\epsilon)$-performative omnipredictor that is $\mathcal{W}$-universally adaptable over $\mathcal{D}$.*

Proof. The proposition follows as a corollary of Theorem 5.2.5. The key observation is that multiplying by the importance weight $\omega(x)$ allows us to switch from an expectation over $\mathcal{D}$ to an expectation over $\mathcal{D}_\omega$. By the definition of $\mathcal{D}_\omega$, we have that for supported $x \in \mathcal{X}$, ω is the odds ratio,

$$\omega(x) = \frac{\mathcal{D}_\omega(x)}{\mathcal{D}(x)}.$$

Further, by the definition of ℓ_ω, for any $h : \mathcal{X} \to \hat{\mathcal{Y}}$ and any outcome probability model p, the following equality of expectations holds

$$\mathbb{E}_{\substack{x\sim\mathcal{D} \\ y\sim p(x,h(x))}}[\ell_\omega(x, h(x), y)] = \mathbb{E}_{\substack{x\sim\mathcal{D} \\ y\sim p(x,h(x))}}[\omega(x) \cdot \ell(x, h(x), y)] \tag{5.2}$$

$$= \mathbb{E}_{\substack{x\sim\mathcal{D}_\omega \\ y\sim p(x,h(x))}}[\ell(x, h(x), y)], \tag{5.3}$$

where we rely on the identity that for any function $g : \mathcal{X} \to \mathbb{R}$,

$$\mathbb{E}_{\mathcal{D}}[g(x) \cdot \omega(x)] = \mathbb{E}_{\mathcal{D}}[g(x) \cdot \mathcal{D}_\omega(x)/\mathcal{D}(x)] = \mathbb{E}_{\mathcal{D}_\omega}[g(x)].$$

The equality in Equation 5.2 immediately implies that if $\tilde{p}$ is $(\mathcal{L}_{\mathcal{W}}, \mathcal{H}, \epsilon)$-POI over $\mathcal{D}$, then $\tilde{p}$ is $(\mathcal{L}, \mathcal{H}, \epsilon)$-POI over every $\mathcal{D}_\omega \in \mathcal{D}_{\mathcal{W}}$. That is, by applying the identity to the expectation under Nature's outcomes $y^* \sim p^*(x, h(x))$ and separately to the expectation under the modeled outcomes $\tilde{y} \sim \tilde{p}(x, h(x))$, $(\mathcal{L}_{\mathcal{W}}, \mathcal{H}, \epsilon)$-performative OI imples that we obtain indistinguishability for all $\ell \in \mathcal{L}$, $h \in \mathcal{H}$ and $\mathcal{D}_\omega \in \mathcal{D}_{\mathcal{W}}$:

$$\left|\mathbb{E}_{\substack{x\sim\mathcal{D}_\omega \\ y^*\sim p^*(x,h(x))}}[\ell(x, h(x), y^*)] - \mathbb{E}_{\substack{x\sim\mathcal{D}_\omega \\ \tilde{y}\sim \tilde{p}(x,h(x))}}[\ell(x, h(x), \tilde{y})]\right| \leq \epsilon.$$

The corresponding statement for performative decision OI is a bit more subtle. Whereas above, the decision rules $h \in \mathcal{H}$ do not depend in any way on ω, the optimal decision rule $\tilde{f}_{\ell_\omega}$ based on $\tilde{p}$, is allowed to depend on the specified loss and, thus, on ω. Still, we argue that for any ω, $\tilde{f}_{\ell_\omega} = \tilde{f}_\ell$. This equality follows by the fact that the optimal decision rule is chosen pointwise, for each $x \in \mathcal{X}$. In particular, for all $x \in \mathcal{X}$, scaling the loss by $\omega(x)$ changes the scale of the optimization, but not the minimizer:

$$\tilde{f}_{\ell_\omega}(x) = \arg\min_{\hat{y}\in\hat{\mathcal{Y}}} \mathbb{E}_{y\sim\tilde{p}(x,\hat{y})}[\omega(x) \cdot \ell(\hat{y}, y)] = \arg\min_{\hat{y}\in\hat{\mathcal{Y}}} \mathbb{E}_{y\sim\tilde{p}(x,\hat{y})}[\ell(\hat{y}, y)] = \tilde{f}_\ell(x).$$

Thus, the same identities from above can be applied to prove that if $\tilde{p}$ is $(\mathcal{L}_\omega, \epsilon)$-DOI for a fixed distribution $\mathcal{D}$, then it is also $(\mathcal{L}, \epsilon)$-DOI for every $\mathcal{D}_\omega \in \mathcal{D}_{\mathcal{W}}$. The proposition follows by applying Theorem 5.2.5 separately over each $\mathcal{D}_\omega \in \mathcal{D}_{\mathcal{W}}$. □

Before moving on, we highlight that designing omnipredictors requires the learner to account for possible shifts in the distribution at *training* time, not at test time. At test time, the learner simply chooses predictions $\hat{y}$ according to the function $\tilde{f}_\ell$, without needing to first infer what the underlying distribution $\mathcal{D}$ may be. The decision rule $\tilde{f}_\ell$ is simultaneously optimal for all of them. This design choice shifts the burden of technical sophistication and expertise from the user of the system to its designer. The user is free to focus on the choice of loss function ℓ to balance between forecasting and steering knowing that naive usage of $\tilde{f}_\ell$ is guaranteed to work.

5.4 Learning Algorithms for Performative Omniprediction

In this section, we introduce a general purpose algorithm, POI-Boost, which provably returns a performative omnipredictor $\tilde{p}$ for any class of hypothesis $\mathcal{H}$ and collection of losses $\mathcal{L}$. Our algorithmic approach is centered on establishing two reductions. First, we prove that, similar to previous work in the OI literature, learning Performative OI predictors reduces to the problem of *auditing* for outcome indistinguishability.

The auditing problems we reduce to involve determining whether the losses under the decision rules in $\mathcal{H}$ and $\{\tilde{f}_\ell\}$ are the same for Nature's outcomes and our modeled outcomes *under outcome performativity*. While in the supervised learning setting we only need to reason about a single outcome distribution, in our setting, different different decision rules induce different distributions over outcomes, and we want to audit for indistinguishability over each of these induced distributions. Despite this challenge, we show that, given access to appropriately randomized data, we can reduce this performative auditing problem to standard supervised learning primitives. In this second reduction, we make use of computational and statistical assumptions: access to an appropriate supervised learner (computational) and access to randomized control data (statistical).

While we instantiate our algorithm with specific computational and statistical assumptions, the framework for learning is completely generic and modular. In particular, any solution to the auditing problem can be used to implement the algorithm. It stands to reason that our assumptions could be relaxed in the future, or that in certain settings, incomparable assumptions lead to more effective auditing, which in turn would lead to more efficient learning of performative omnipredictors.

Performative OI Boost (POI-Boost)

Input: Set of losses $\mathcal{L}$, hypotheses $\mathcal{H}$, distribution $\mathcal{D}$, tolerance $\epsilon > 0$

Initialize: $q^{(1)}(\cdot) \leftarrow [1/2, \ldots, 1/2] \in [0,1]^{|\hat{\mathcal{Y}}|}$

For $t = 1, 2, \ldots$

- **If:** there exists **a**) $(h, \ell) \in \mathcal{H} \times \mathcal{L}$ or, **b**) $(h, \ell) \in \{(f_{\ell,t}, \ell) : \ell \in \mathcal{L}\}$

$$\underbrace{|\mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ \tilde{y}\sim\tilde{p}(x,h(x))}}[\ell(x,h(x),\tilde{y})] - \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,h(x))}}[\ell(x,h(x),y^*)]|}_{\mathsf{err}_t} \geq \epsilon \tag{5.4}$$

  Then: Update the representation $q^{(t)}$:

$$q^{(t+1)}(\cdot) \leftarrow \Pi\left(q^{(t)}(\cdot) - \eta^{(t)}\, v_{\ell,h}(\cdot) \right) \tag{5.5}$$

  where $\eta_t = -\epsilon \cdot \mathsf{sign}(\mathsf{err}_t) \;/\; \ell_{\max}$ and

$$v_{\ell,h}(\cdot) = \begin{bmatrix} (\ell(\cdot,\hat{y}_1,1) - \ell(\cdot,\hat{y}_1,0))\; \mathbf{1}\{h(\cdot)=\hat{y}_1\} \\ \ldots \\ (\ell(\cdot,\hat{y}_k,1) - \ell(\cdot,\hat{y}_k,0))\; \mathbf{1}\{h(\cdot)=\hat{y}_k\} \end{bmatrix} \in \mathbb{R}^{|\hat{\mathcal{Y}}|}, \quad \hat{\mathcal{Y}} = \{\hat{y}_1, \ldots, \hat{y}_k\} \tag{5.6}$$

- **Else:** terminate and return the function $\tilde{p}(x, \hat{y}) = q^{(t)}(x)[\ \hat{y}\ ]$

Figure 5.2: Algorithm for generating performative omnipredictors. The algorithm proceeds by repeatedly verifying whether the intermediate predictors $p^{(t)}$ satisfy the POI definition, outlined in **a**), as well as the DOI definition, outlined in **b**). If neither is violated, the procedure terminates. Otherwise, the algorithm implicitly updates the representation $q^{(t)}$ of the predictor $p^{(t)}$. Given an input x, $q^{(t)}(x)$ is a vector of length $|\hat{\mathcal{Y}}|$ whose $\hat{y}$ entry, $q^{(t)}(x)[\ \hat{y}\ ]$, represents $p^{(t)}(x, \hat{y})$. The operator Π clips entries of its input vector to lie in $[0, 1]$. For the sake of clarity, here we present the simplest version of the algorithm where the search outlined in Equation 5.4 is proper, however this condition can be easily relaxed as discussed in Section 5.4.

Reducing Indistinguishability to Auditing

We start by establishing our first reduction. We prove that the POI-Boost algorithm (Figure 5.2) returns a performative omnipredictor $\tilde{p}$ after a small, polynomial number

of calls to an auditing subroutine (described in Equation 5.4 & Figure 5.3), without yet describing the runtime or sample complexity of the auditing step itself. We address these questions in the next subsection. The algorithm works for any outcome performative problem where the number of predicted labels $\hat{\mathcal{Y}}$ is finite and the loss functions are bounded.

Representing Predictors. For the sake of our analysis, it is helpful to distinguish between the predictor $\tilde{p} : \mathcal{X} \times \hat{\mathcal{Y}} \to [0,1]$ as a function, and the implementation of $\tilde{p}$ in code. In our learning algorithm, we represent the function $\tilde{p}$ in terms of vector-valued functions $\tilde{q} : \mathcal{X} \to [0,1]^{|\hat{\mathcal{Y}}|}$. Given $x \in \mathcal{X}$, $\tilde{q}(x)$ is a vector of length $|\hat{\mathcal{Y}}|$ whose $\hat{y}$ entry, $q^{(t)}(x)[\ \hat{y}\ ]$, represents $\tilde{p}(x, \hat{y})$.

Of course, there is a correspondence between these functions $\tilde{p}$ and $\tilde{q}$ where each $\tilde{p}$ leads to a unique $\tilde{q}$ and vice versa. The key difference is that $\tilde{q}(x)$ returns $\tilde{p}(x, \hat{y})$ for all $\hat{y} \in \hat{\mathcal{Y}}$ in a single function call, while computing the same information using the direct $\tilde{p}$ representation would require $|\hat{\mathcal{Y}}|$ functions calls. While this might seem like a minor detail, these representations have meaningful differences in terms of the circuit complexity of performative omnipredictors. Crucially, for any loss ℓ, to compute $\tilde{f}_\ell(x)$, we need the value of of $\tilde{p}(x, \hat{y})$ for all $\hat{y} \in \hat{\mathcal{Y}}$, and this computation can be performed using a single call to $\tilde{q}$. In other words, we perform $|\hat{\mathcal{Y}}|$ times the work per call to $\tilde{q}$ to avoid $|\hat{\mathcal{Y}}|$ *recursive* calls to the predictor within the construction. Avoiding further calls to these functions avoids branching factors and an exponential blowup in the complexity of the resulting predictors as we illustrate.

Algorithm Description. As outlined in Figure 5.2, POI-Boost is an iterative algorithm which non-parametrically constructs a predictor $\tilde{p}$, represented in terms of a vector-valued function $\tilde{q}$, by stringing together copies of circuits which compute losses ℓ and decision rules h. At each iteration, the algorithm first appeals to auditing subroutines to check if there: is a) a pair h, l for which the current predictor $p^{(t)}$, fails the performative OI guarantee, or b) a loss function ℓ for which the decision rule $f_{\ell,t}$ fails the decision OI guarantee. If neither condition is violated, then the algorithm terminates since $p^{(t)}$ satisfies both indistinguishability conditions and consequently must be an omnipredictor as per Theorem 5.2.5.

On the other hand, if one of these conditions is violated, we perform an update to the representation $q^{(t)}$ of the current predictor $p^{(t)}$. These updates nudge the predictor closer to p^* by essentially performing gradient descent in function space [54]. These updates are done implicitly in the sense that we can update the representation $q^{(t)}$ for all x in $\mathcal{X}$ simultaneously by simply adding a copy of the circuit computing $v_{\ell,h} : \mathcal{X} \to \mathbb{R}^{|\hat{\mathcal{Y}}|}$

(Equation 5.6) which is defined in terms of a loss ℓ and decision rule h. By bounding the total number of updates via a potential argument, we can ensure that we don't add too many copies of these functions so that the final predictor is computationally efficient.

Proposition 5.4.1. *The POI-Boost algorithm described in Figure 1 terminates in at most* $|\hat{\mathcal{Y}}|\ell_{\max}^2/\epsilon^2$ *many iterations and returns a predictor* $\tilde{p}$ *that is* $(\mathcal{L}, \mathcal{H}, \epsilon)$*-performative OI and* $(\mathcal{L}, \epsilon)$*-performative decision OI. Consequently,* $\tilde{p}$ *is a* $(\mathcal{L}, \mathcal{H}, 2\epsilon)$*-performative omnipredictor.*

Proof. The guarantee that $\tilde{p}$ is performative decision OI and performative OI follow directly from the termination criterion. Therefore, the proposition follows from proving that this termination criteria is met within the stated number of iterations.

The key insight is that if the indistinguishability constraint in Equation 5.4 is violated for any ℓ or h, then updating the representation $q^{(t)}$ ensures that we will have made nontrivial progress on a common potential function. Since this potential is bounded from above and below, and we make nontrivial progress with every update, the total number of updates must be bounded. In more detail, first, note that for any model p, loss ℓ, hypothesis h, and $\mathcal{Y} = \{0, 1\}$:

$$\begin{aligned}\mathbb{E}_{\substack{x\sim\mathcal{D}\\ y\sim p(x,h(x))}}[\ell(x,h(x),y)] &= \mathbb{E}_x\mathbb{E}_{y|x}[\ell(x,h(x),y)] \\ &= \mathbb{E}_x[\ell(x,h(x),0) + (\ell(x,h(x),1) - \ell(x,h(x),0)) \cdot p(x,h(x))].\end{aligned}$$

From this rewriting, and the definition of $v_{\ell,h}$ in Equation 5.6, the difference in performative risks between the predictors $p^{(t)}$ and p^* for a pair ℓ, h can be expanded as,

$$\mathbb{E}_{\substack{x\sim\mathcal{D}\\ y_t\sim p^{(t)}(x,h(x))}}[\ell(x,h(x),y_t)] - \mathbb{E}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,h(x))}}[\ell(x,h(x),y^*)] \tag{5.7}$$

$$= \mathbb{E}_x\langle q^{(t)}(x) - q^*(x),\ v_{\ell,h}(x)\rangle. \tag{5.8}$$

Now consider the potential, written in terms of the representations $q^{(t)}$,

$$\mathbb{E}_x\|q^{(t+1)}(x) - q^*(x)\|^2.$$

By definition of the update rule in the algorithm, this potential is equal to:

$$\mathbb{E}_x\|\Pi\left(q^{(t)}(x) - \eta^{(t)}v_{\ell,h}(x)\right) - q^*(x)\|^2.$$

Because the projection (or clipping) operator Π can only decrease the distance to p^*, if an update is performed, the difference between potentials at adjacent time steps,

$$\mathbb{E}_x\|q^{(t+1)}(x) - q^*(x)\|^2 - \mathbb{E}_x\|q^{(t)}(x) - q^*(x)\|^2,$$

is upper bounded by the sum of two terms,

$$-2\eta_t \mathbb{E}_x \langle q^{(t)}(x) - q^*(x), v_{\ell,h}(x) \rangle + \eta_t^2 \mathbb{E}_x \|v_{\ell,h}(x)\|^2.$$

Using the identity from Equation 5.8 and the definition of $v_{\ell,h}$ from Equation 5.6, this is equal to:

$$-2\eta_t \mathsf{err}_t + \eta_t^2 \mathbb{E}_x[(\ell(x, h(x), 1) - \ell(x, h(x), 0))^2].$$

Because losses lie in $[0, \ell_{\max}]$, the second term is less than $\ell_{\max}^2 \eta_t^2$. Furthermore, from the auditing guarantee, $|\mathsf{err}_t| > \epsilon$. By setting the step size $\eta^{(t)}$ to be $-\epsilon \cdot \mathrm{sign}(\mathsf{err}_t) \;/\; \ell_{\max}$, we conclude that the difference in potentials across adjacent time steps satisfies,

$$2\eta_t \mathsf{err}_t + \eta_t^2 \mathbb{E}_x[(\ell(x, h(x), 1) - \ell(x, h(x), 0))^2] \leq -2\eta_t \mathsf{err}_t + \eta_t^2 \ell_{\max}^2 \leq -\epsilon^2 / \ell_{\max}^2.$$

Since the potential is nonnegative and bounded above by $|\hat{\mathcal{Y}}|$, the maximum number of iterations until the termination criterion is met must be at most $|\hat{\mathcal{Y}}| \ell_{\max}^2 / \epsilon^2$. □

An important consequence of this result is that it reveals the existence of omnipredictors $\tilde{p}$ that admit computationally efficient approximations. This result is subtle, even in light of previous work on OI-style boosting algorithms. Intuitively, the final $\tilde{p}$ is built out by stringing together copies of functions in $\mathcal{H}$, $\mathcal{L}$, and decision rules $f_{\ell,t}$. Because these decision rules $f_{\ell,t}$ are defined in terms of an optimization procedure involving the intermediate constructions $p^{(t)}$, which themselves depend on previous models $p^{(t-1)}$, a naive implementation of $\tilde{p}$ can result in a recursion that induces an exponentially large circuit. Specifically, the naive implementation would make $|\hat{\mathcal{Y}}|$ recursive calls to the prior circuit in order to compute $f_{\ell,t}$, resulting in a growth rate of $|\hat{\mathcal{Y}}|^t$.

However, by carefully ordering the relevant computations and "caching" previous work, we avoid this blow-up. The key insight is the following. By designing a circuit that computes the value of $\tilde{p}(x, \hat{y})$ for every $\hat{y} \in \hat{\mathcal{Y}}$ simultaneously, we can avoid recursive calls to the circuit. By maintaining the intermediate computations of $q^{(t)}$, we can avoid a branching factor in the program and preserve efficiency.

Theorem 5.4.2. *Assume that the functions in $\mathcal{H}$ and $\mathcal{L}$ are computable by circuits of size at most s, then the predictor $\tilde{p}$ returned by the POI-Boost algorithm has size at most $\ell_{\max}^2/\epsilon^2 \cdot \mathrm{poly}(s, |\hat{\mathcal{Y}}|)$.*

Proof. The final predictor consists of a summation of the initial prediction, followed by the update from each iteration. We bound the growth of the circuit computing the predictor by induction. Formally, let S_t be the circuit size for computing $q^{(t)}$. Then, we show that $S_{t+1} \leq S_t + \mathrm{poly}(|\hat{\mathcal{Y}}|, s)$ for all $t \geq 1$. Thus, by the overall bound on the iteration complexity, the final predictor can be implemented using a circuit of size $S \leq \ell_{\max}^2/\epsilon^2 \cdot \mathrm{poly}(s, |\hat{\mathcal{Y}}|)$. To

begin, the initial constant predictor $q^{(1)}$ can be implemented using a circuit of size at most $S_1 = |\hat{\mathcal{Y}}| \leq \text{poly}(s, |\hat{\mathcal{Y}}|)$ by hard-coding the constant vector.

By the update rule, each update incorporates a function of the form,

$$g^{(t)}(x) := \eta^{(t)} v^{(t)}_{\ell,h}(x) = \eta^{(t)} \begin{bmatrix} (\ell^{(t)}(x, \hat{y}_1, 1) - \ell^{(t)}(x, \hat{y}_1, 0))\, \mathbf{1}\{h^{(t)}(x) = \hat{y}_1\} \\ \dots \\ (\ell^{(t)}(x, \hat{y}_k, 1) - \ell^{(t)}(x, \hat{y}_k, 0))\, \mathbf{1}\{h^{(t)}(x) = \hat{y}_k\} \end{bmatrix} \in \mathbb{R}^{|\hat{\mathcal{Y}}|}, \quad (5.9)$$

where $\ell^{(t)}$ and $h^{(t)}$ define the test function surfaced by the auditing subroutine at time step t. Within these updates, the function $h^{(t)}$ may *a*) come from $\mathcal{H}$ due to a POI violation or *b*) equal $f_{\ell,t}$ for some $\ell \in \mathcal{L}$ due to a DOI violation.

In the first case where $h^{(t+1)} \in \mathcal{H}$, $q^{(t+1)}(x)$ can be computed by evaluating evaluating $q^{(t)}(x)$, and then evaluating $h(x)$ and $\ell(x, \hat{y}, y)$, for every $\hat{y}$ and y. By assumption, the latter operations require circuits of size at most $\text{poly}(s, |\hat{\mathcal{Y}}|)$. Paired with the inductive hypothesis, the resulting circuit size can be bounded as $S_{t+1} \leq S_t + \text{poly}(s, |\hat{\mathcal{Y}}|) \leq (t+1) \cdot \text{poly}(s, |\hat{\mathcal{Y}}|)$.

For the second case, we recall the definition of $f_{\ell,t}(\cdot)$, we can express its computation as a minimization over $\hat{\mathcal{Y}}$ of expected losses that depend on $q^{(t)}(\cdot)[\hat{y}]$ for each $\hat{y} \in \hat{\mathcal{Y}}$.

$$f_{\ell,t}(x) = \arg\min_{\hat{y} \in \hat{\mathcal{Y}}} \{\ell(x, \hat{y}, 0) + (\ell(x, \hat{y}, 1) - \ell(x, \hat{y}, 0)) \cdot q^{(t)}(x)[\hat{y}]\}.$$

Importantly, to compute each term in the minimization, we only need to compute the vector $q^{(t)}(x)$ once. The remaining terms, $\ell(x, \hat{y}, 0)$ and $\ell(x, \hat{y}, 1)$ (for every $\hat{y}$), can again be computed by a circuit of size $\text{poly}(|\hat{\mathcal{Y}}|, s)$. Since the minimization itself can be done by linearly enumerating over $\hat{\mathcal{Y}}$, we again preserve the invariant that $S_{t+1} \leq S_t + \text{poly}(|\hat{\mathcal{Y}}|, s)$. □

Reducing Auditing to Supervised Learning

Having shown how omniprediction reduces to an auditing problem, we now complete our analysis of the POI-Boost algorithm by showing that auditing itself reduces to cost-sensitive classification over a single, static distribution. In doing so, we address the statistical and computational complexity of solving this auditing step.

From examining the auditing condition in Figure 5.3, perhaps the most obvious strategy is to choose a decision rule h, and to collect a dataset of triples $(x, \hat{y}, y^*)$ where $\hat{y} = h(x)$ for every x and $y^* \sim p^*(x, h(x))$. If the loss ℓ is bounded, a standard application of Hoeffding's inequality shows that empirical risk of the loss concentrates around its expectation:

$$\frac{1}{n} \sum_{i=1}^{n} \ell(x_i, h(x_i), y_i^*) \approx \mathbb{E}_{\substack{x \sim \mathcal{D} \\ y^* \sim p^*(x, h(x))}} [\ell(x, h(x), y_*)].$$

Audit($\mathcal{H}, \mathcal{L}, \tilde{p}, \epsilon$)

If: there exists $(h, \ell) \in \mathcal{H} \times \mathcal{L}$

$$\left|\mathbb{E}_{\substack{x\sim\mathcal{D}\\ \tilde{y}\sim\tilde{p}(x,h(x))}} \ell(x, h(x), \tilde{y}) - \mathbb{E}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,h(x))}} \ell(x, h(x), y^*)\right| \geq \epsilon$$

Then: return h, l

Else: return False

Figure 5.3: The key auditing step in the POI-Boost algorithm. In each iteration of the algorithm, we run two auditing steps: once to check for the POI condition over $\mathcal{H} \times \mathcal{L}$ and once to check for the DOI condition over $\{(f_{\ell,t}, \ell) : \ell \in \mathcal{L}\}$. See the proof of Corollary 5.4.8 for further discussion.

Therefore, one could implement the auditing step by enumerating over all h, deploying h to collect a new dataset every time, and then nonadaptively computing the empirical performative risk of h on every $\ell \in \mathcal{L}$. This procedure would however require $\tilde{O}(|\mathcal{H}|/\epsilon^2 \log |\mathcal{L}|)$ many samples.

On the other hand, if we have access to randomized predictions $\hat{y}$, we can estimate the empirical risk of every pair h, ℓ off of a *single* distribution by using inverse propensity scoring. The following lemma is well-known within various communities, and, in particular, the contextual bandits literature (see e.g. [3, 15]).[7]

We use the shorthand $(x, \hat{y}, y) \sim \mathcal{D}_{\text{rct}}$ to denote the sampling process where inputs are sampled from the base distribution $x \sim \mathcal{D}$, decisions $\hat{y}$ are assigned uniformly at random, $\hat{y} \sim \text{Unif}(\hat{\mathcal{Y}})$, and the outcomes are sampled according to Nature's model $y^* \sim p^*(x, \hat{y})$.

Lemma 5.4.3. *Assume that $\hat{\mathcal{Y}}$ is a finite set. Then, for any hypothesis $h : \mathcal{X} \to \hat{\mathcal{Y}}$,*

$$\mathbb{E}_{\substack{x\sim\mathcal{D}_x\\ y^*\sim p^*(x,h(x))}} [\ell(x, h(x), y^*)] = |\hat{\mathcal{Y}}| \cdot \mathbb{E}_{(x,\hat{y},y^*)\sim\mathcal{D}_{\text{rct}}}[\ell(x, \hat{y}, y^*)\mathbf{1}\{h(x) = \hat{y}\}].$$

Proof. We present the proof for the case where $\hat{\mathcal{Y}} = \{0, 1\}$ is binary, but the general case

[7] This result can be generalized to the case where for every x, $\hat{y} \sim q(x)$ for some known distribution q, where $q \neq \text{Unif}(\hat{\mathcal{Y}})$ but where q has full support over $\hat{\mathcal{Y}}$.

follows the same pattern. We expand out $\mathbb{E}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,h(x))}}[\ell(x,h(x),y^*)]$ as:

$$= \mathbb{E}_{\substack{x\sim\mathcal{D}\\ y^*_{(1)}\sim p^*(x,1), y^*_{(0)}\sim p^*(x,0)}}\left[\ell(x,1,y^*_{(1)})\mathbf{1}\{h(x)=1\} + \ell(x,0,y^*_{(0)})\mathbf{1}\{h(x)=0\}\right]$$
$$= \mathbb{E}_{\substack{x\sim\mathcal{D}\\ y^*_{(1)}\sim p^*(x,1)}}\left[\ell(x,1,y^*_{(1)})\mathbf{1}\{h(x)=1\}\right] + \mathbb{E}_{\substack{x\sim\mathcal{D}\\ y^*_{(0)}\sim p^*(x,0)}}\left[\ell(x,1,y^*_{(0)})\mathbf{1}\{h(x)=0\}\right].$$

Reweighting the term on the right hand side, we observe our desired equality:

$$\mathbb{E}_{\substack{x\sim\mathcal{D}\\ \hat{y}\sim\text{Ber}(1/2)\\ y^*\sim p^*(x,\hat{y})}}[\ell(x,\hat{y},y^*)\mathbf{1}\{h(x)=\hat{y}\}] = \frac{1}{2}\mathbb{E}_{\substack{x\sim\mathcal{D}\\ y^*_{(1)}\sim p^*(x,1)}}\left[\ell(x,1,y^*_{(1)})\mathbf{1}\{h(x)=1\}\right] + \frac{1}{2}\mathbb{E}_{\substack{x\sim\mathcal{D}\\ y^*_{(0)}\sim p^*(x,0)}}\left[\ell(x,1,y^*_{(0)})\mathbf{1}\{h(x)=0\}\right].$$

□

There are two main takeaways from this lemma. First, it shows that the statistical complexity of auditing can be exponentially better than the the naive strategy outlined previously.

Corollary 5.4.4. *Let $\{(x_i,\hat{y}_i,\tilde{y}_i)\}_{i=1}^n$ be a dataset of n i.i.d samples from $\mathcal{D}_{\text{rct}}$. If*

$$n \geq \frac{2\ell_{\max}^2|\hat{\mathcal{Y}}|^2\cdot\log(2|\mathcal{H}||\mathcal{L}|/\delta)}{\epsilon^2},$$

then with probability $1-\delta$,

$$\max_{h\in\mathcal{H},\ell\in\mathcal{L}}\left|\mathbb{E}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,h(x))}}[\ell(x,h(x),y^*)] - \frac{1}{n}\sum_{i=1}^n|\hat{\mathcal{Y}}|\cdot\ell(x_i,\hat{y}_i,y_i^*)\mathbf{1}\{h(x_i)=\hat{y}_i\}\right| \leq \epsilon.$$

Proof. From the previous lemma, we have that for any loss ℓ and decision rule h,

$$\mathbb{E}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,h(x))}}[\ell(x,h(x),y^*)] = |\hat{\mathcal{Y}}|\cdot\mathbb{E}_{(x,\hat{y},y^*)\sim\mathcal{D}_{\text{rct}}}[\ell(x,\hat{y},y^*)\mathbf{1}\{h(x)=\hat{y}\}].$$

Because $|\hat{\mathcal{Y}}|\cdot\ell(x_i,\hat{y}_i,y_i^*)\mathbf{1}\{h(x_i)=\hat{y}_i\}$ is uniformly bounded by $\ell_{\max}|\hat{\mathcal{Y}}|$, we can apply Hoeffding's inequality to argue that the probability that the empirical estimate is far from the true expectation

$$\left|\mathbb{E}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,h(x))}}[\ell(x,h(x),y^*)] - \frac{1}{n}\sum_{i=1}^n|\hat{\mathcal{Y}}|\cdot\ell(x_i,\hat{y}_i,y_i^*)\mathbf{1}\{h(x_i)=\hat{y}_i\}\right| > \epsilon$$

is bounded as $2\exp\left(\frac{2\epsilon^2}{n(\ell_{\max}|\hat{\mathcal{Y}}|)^2}\right)$. The result follows by rearranging for failure probability δ, and taking a union bound over all $h\in\mathcal{H}$ and $\ell\in\mathcal{L}$. □

Consequently, for a single iteration of the POI-Boost algorithm, the auditing step can be implemented by enumerating over all $\mathcal{L}$ and $\mathcal{H}$ and non-adaptively evaluating their empirical risks on a single dataset of size $\tilde{O}(\ell_{\max}^2|\hat{\mathcal{Y}}|^2/\epsilon^2 \log(|\mathcal{H}||\mathcal{L}|))$.[8] Typically, we think of $|\hat{\mathcal{Y}}|$ as a small constant and the class of decision rules $\mathcal{H}$ as a rich collection. From this result, we see that at least statistically, we can hope to design omnipredictors that are optimal with respect to an exponential number of losses and decision rules.

Here, we present the simplest possible analysis of this result and state our bounds for finite classes $\mathcal{H}$ and $\mathcal{L}$. It is certainly feasible to achieve sharper results and to state bounds in terms of VC-dimension or other sharper notions of statistical complexity. However, the goal of our initial work on outcome performativity is not to establish the tightest bounds, but to provide a broad overview of what is possible. We hope future work will provide a precise understanding of the sample complexity of omniprediction in outcome performativity.

The following proposition summarizes the sample complexity of omniprediction if the auditing steps for the POI and DOI conditions are implemented via a naive learner that linearly enumerates over all h, ℓ and evaluates their empirical risk on a single dataset of RCT samples.

Proposition 5.4.5. *Given labeled data $(x, \hat{y}, y^*) \sim \mathcal{D}_{\text{rct}}$ drawn from Nature and unlabeled samples $x \sim \mathcal{D}$, the POI-boost can be implemented using at most:*

- $O(\ell_{\max}^2|\hat{\mathcal{Y}}|^2 \log(\frac{|\mathcal{H}||\mathcal{L}|}{\delta})/\epsilon^2 + \ell_{\max}^4|\hat{\mathcal{Y}}|^3 \log(\frac{|\mathcal{L}|\ell_{\max}|\hat{\mathcal{Y}}|}{\delta\epsilon})/\epsilon^4)$ *labeled samples*
- $O(\ell_{\max}^4|\hat{\mathcal{Y}}|^3 \log(\frac{|\mathcal{H}||\mathcal{L}|\ell_{\max}|\hat{\mathcal{Y}}|}{\delta\epsilon})/\epsilon^4)$ *unlabeled samples*

Proof. In each iteration of the POI-boost algorithm, we need to audit for the POI and DOI guarantees (conditions *a* and *b*). We can implement each of the auditing steps by explicit enumeration.

For POI, at each iteration t we enumerate over $\mathcal{H}$ and $\mathcal{L}$ and evaluate the empirical counterparts of

$$\mathbb{E}_{\substack{x\sim\mathcal{D}\\ \tilde{y}\sim p^{(t)}(x,h(x))}}[\ell(x,h(x),\tilde{y})] \quad \text{and } \mathbb{E}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,h(x))}}[\ell(x,h(x),y^*)]. \tag{5.10}$$

By Corollary 5.4.4, the empirical versions of these quantities concentrate around their expectations. To get an ϵ approximation, with probability $1-\delta$, we require at most $O(\ell_{\max}^2|\hat{\mathcal{Y}}|^2 \log(|\mathcal{H}||\mathcal{L}|/\delta)/\epsilon^2)$ many samples. At each iteration t, the expectation on the left changes, since we update $p^{(t)}$. However, to evaluate this expectation we only need

[8] An analogous result applies if we replace p^* by $\tilde{p}$, which we assume that the learner can easily sample from.

unlabeled samples, since labels $\tilde{y}$ come from our own model $p^{(t)}$. On the other hand, the expectation on the right in Equation 5.10 does not depend on t, so we need not recompute it at every iteration. Because the total number of iterations is bounded by $\ell_{\max}^2|\hat{\mathcal{Y}}|/\epsilon^2$, applying a union bound on δ, to achieve the POI guarantee we only need a total of $O(\ell_{\max}^4|\hat{\mathcal{Y}}|^3\log(|\mathcal{H}||\mathcal{L}|\ell_{\max}|\hat{\mathcal{Y}}|\epsilon^{-1}\delta^{-1})/\epsilon^4)$ unlabeled samples and

$$O(\ell_{\max}^2|\hat{\mathcal{Y}}|\log(|\mathcal{H}||\mathcal{L}|/\delta)/\epsilon^2)$$

labeled samples.

For the DOI guarantee outline in condition b, we instead need to approximate

$$\mathbb{E}_{\substack{x\sim\mathcal{D}\\ \tilde{y}\sim p^{(t)}(x,f_{\ell,t}(x))}}[\ell(x,f_{\ell,t}(x),\tilde{y})] \quad \text{and } \mathbb{E}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,f_{\ell,t}(x))}}[\ell(x,f_{\ell,t}(x),y^*)]. \tag{5.11}$$

Note that both of these expectations now depend on t, because the decision rules $f_{\ell,t}$ can change between iterations. Again, by Corollary 5.4.4, if we enumerate over all $|\mathcal{L}|$ losses and decision rules $f_{\ell,t}$ at each iteration, the empirical counterparts of these expressions on a dataset of size $O(\ell_{\max}^2|\hat{\mathcal{Y}}|^2\log(|\mathcal{L}|/\delta)/\epsilon^2)$ concentrates. Collecting a new dataset at every iteration, we get that the total number of labeled (and unlabeled) samples is bounded by $O(\ell_{\max}^4|\hat{\mathcal{Y}}|^3\log(|\mathcal{L}||\hat{\mathcal{Y}}|\ell_{\max}\delta^{-1}\epsilon^{-1})/\epsilon^4)$. □

Cost-Sensitive Classification. The second main takeaway from Lemma 5.4.3 is that auditing can now be rewritten as as the solution to a cost-sensitive multiclass classification problem over $|\hat{\mathcal{Y}}|$ many classes. This result completes our analysis showing how omniprediction can be reduced to basic supervised learning problems.

In light of previous results, the main benefit of this reduction is that it enabled the design of oracle-efficient algorithms which can be faster than the naive learner used in Proposition 5.4.5. We start by first defining what we mean by cost-sensitive classification.

Definition 5.4.6. Let $\mathcal{X}$ be a feature space, $\hat{\mathcal{Y}}$ be a finite set of k classes, and $\mathcal{D}$ be a distribution over $\mathcal{X}\times[-1,1]^k$. For $(x,c)\sim\mathcal{D}$, we say that c is a cost vector whose entries $c(\hat{y})$ denote the costs of predicting label $\hat{y}$ on feature x. An algorithm $\mathcal{A}_{\text{csc}}$ is an ρ-cost-sensitive learner for a hypothesis class $\mathcal{H}$ if for any distribution $\mathcal{D}$ over $\mathcal{X}\times[-1,1]^k$, promised that there exists $h\in\mathcal{H}$ such that $\mathbb{E}_{(x,c)\sim\mathcal{D}}c(h(x))\le-\rho$, $\mathcal{A}_{\text{csc}}$ returns a hypothesis h' such that $\mathbb{E}_{(x,c)\sim\mathcal{D}}c(h'(x))\le-\rho/2$.

Cost-sensitive classification is a well-studied supervised learning problem for which many, both passive and active learning algorithms, have been designed [1, 9, 20, 45, 46]. There are a number of software packages that can be used to solve applied cost-senstive classification problems [68]. Like many problems in computational learning theory, cost-sensitive classification is known to be hard in the worst-case, but can be solved effectively

in practice. As such, our goal is to design *oracle-efficient* learning algorithms, that make an small number of calls to cost-sensitive learner.

Here, we frame a "weak" version of the problem where the learning need not be exact, but where the search is proper, in the sense that $\mathcal{A}_{\text{csc}}$ returns a hypothesis in $\mathcal{H}$. This latter condition can easily also be relaxed without changing the overall results. However, we opt to keep it as is for the sake of simplifying the presentation. The following proposition completes our reduction of auditing to supervised learning.

Proposition 5.4.7. *Let $\mathcal{A}_{\text{csc}}$ be a cost-sensitive learner as per Definition 5.4.6. Then, given access to RCT samples $(x, \hat{y}, y^*) \sim \mathcal{D}_{\text{rct}}$, we can solve the auditing problem outlined in Figure 5.3 using $2|\mathcal{L}|$ many calls to $\mathcal{A}_{\text{csc}}$ with parameters $\rho = \epsilon/(4\ell_{\max}|\hat{\mathcal{Y}}|)$.*

Proof. By Lemma 5.4.3 we have that the difference in performative risk between p^* and $\tilde{p}$,

$$\mathbb{E}_{\substack{x\sim\mathcal{D}\\ \tilde{y}\sim\tilde{p}(x,h(x))}}[\ell(x,h(x),\tilde{y})] - \mathbb{E}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,h(x))}}[\ell(x,h(x),y^*)],$$

is equal to:

$$|\hat{\mathcal{Y}}| \cdot \mathbb{E}_{\substack{x\sim\mathcal{D}\\ \hat{y}\sim\text{Unif}(\hat{\mathcal{Y}})\\ \tilde{y}\sim\tilde{p}(x,\hat{y}),y^*\sim p^*(x,\hat{y})}} [\mathbf{1}\{h(x)=\hat{y}\}(\ell(x,h(x),\tilde{y}) - \ell(x,h(x),y^*))] .$$

Now, we note that terms inside the expectation can be written as entries in a cost vector c where for every sample $(x, \hat{y}, y^*, \tilde{y})$ we define the corresponding vector c to be,

$$c_\sigma(h(x)) = \begin{cases} \sigma \cdot (\ell(x,h(x),\tilde{y}) - \ell(x,h(x),y^*)) & \text{if } h(x)=\hat{y} \\ 0 & \text{o.w} \end{cases},$$

Here, $\sigma \in \{\pm 1\}$ and we set $\sigma = 1$ to get the desired equality. Hence, for a fixed loss ℓ, we can transform RCT samples, $x \sim \mathcal{D}, \hat{y} \sim \text{Unif}(\hat{\mathcal{Y}}), y^* \sim p^*(x,\hat{y})$ to a cost sensitive classification problem such that for every $h \in \mathcal{H}$,

$$\mathbb{E}_{x\sim\mathcal{D}'}[c_{+1}(h(x))] = |\hat{\mathcal{Y}}|\mathbb{E}_{\substack{x\sim\mathcal{D}\\ \hat{y}\sim\text{Unif}(\hat{\mathcal{Y}})\\ \tilde{y}\sim\tilde{p}(x,\hat{y}),y^*\sim p^*(x,\hat{y})}} [\mathbf{1}\{h(x)=\hat{y}\}(\ell(x,h(x),\tilde{y}) - \ell(x,h(x),y^*))].$$

To solve the auditing problem outlined in Figure 5.3, we need to check whether the absolute value of the difference is larger than ϵ. To do this, it therefore to suffices to run $\mathcal{A}_{\text{csc}}$ twice (once with $\sigma = 1$ and once with $\sigma = -1$) for every loss $\ell \in \mathcal{L}$ to check if there exists a decision rule $h \in \mathcal{H}$ such that:

$$\mathbb{E}_{x\sim\mathcal{D}'}[c_{+1}(h(x))] \le -\epsilon \text{ or } \mathbb{E}_{x\sim\mathcal{D}'}[c_{-1}(h(x))] \le -\epsilon.$$

Because we normalize the cost vectors to have entries in $[-1,1]$ in Definition 5.4.6, we can scale the vectors c_σ by $1/(4|\hat{\mathcal{Y}}|\ell_{\max})$ and divide the tolerance parameter ϵ by the corresponding amount to match the desired interface. □

End-to-End Analysis

Having now presented these reductions showing how omniprediction can be reduced to cost sensitive classification, we now summarize our results so far and establish end-to-end bounds on the runtime and sample complexities of achieving omniprediction.

Corollary 5.4.8. *Assume that $h \in \mathcal{H}$ and $\ell \in \mathcal{L}$ can be evaluated in time* $\text{poly}(\log(|\mathcal{H}|))$ *and* $\text{poly}(\log(|\mathcal{L}|))$*, respectively. Let $\mathcal{A}_{\text{csc}}$ be a ρ-cost-sensitive weak learner for $\mathcal{H}$ as per Definition 5.4.6. Assume that for any distribution $\mathcal{D}_{\text{csc}}$ over pairs $(x, c) \in \mathcal{X} \times [-1, 1]^k$, $\mathcal{A}_{\text{csc}}$ runs in time* $\text{poly}(\log(|\mathcal{H}|), 1/\rho)$ *and uses at most* $\text{poly}(\log(|\mathcal{H}|), 1/\rho)$ *many samples drawn from $\mathcal{D}_{\text{csc}}$.[9] If the learner has access to samples drawn according $(x, \hat{y}, y^*) \sim \mathcal{D}_{\text{rct}}$, then, the POI-Boost algorithm:*

- *runs in time* $O\left(|\mathcal{L}| \cdot \text{poly}\left(1/\epsilon,\ \ell_{\max},\ |\hat{\mathcal{Y}}|,\ \log|\mathcal{L}|,\ \log|\mathcal{H}|\right)\right)$
- *uses at most* $O\left(\text{poly}\left(1/\epsilon,\ \ell_{\max},\ |\hat{\mathcal{Y}}|,\ \log|\mathcal{L}|,\ \log|\mathcal{H}|\right)\right)$ *many samples*

Proof. To bound the runtime, we note that by Proposition 5.4.1, the maximum number of iterations for POI-Boost is at most $\ell_{\max}^2|\hat{\mathcal{Y}}|/\epsilon^2$. In each iteration, we solve the auditing via two subroutines. One to check for the POI guarantee (condition *a* in Figure 5.2) and another routine to check the DOI guarantee (condition *b* in Figure 5.2). To audit for the POI guarantee, we call the cost-sensitive learner $2|\mathcal{L}|$ times with parameters $\rho = \epsilon/(4\ell_{\max}|\hat{\mathcal{Y}}|)$ as per Proposition 5.4.7, using labels derived from calculating each $\ell \in \mathcal{L}$ and evaluating $p^{(t)}(x, \hat{y})$ in at most $\text{poly}(\log(|\mathcal{L}|)) + \text{poly}(\log(|\mathcal{H}|), 1/\epsilon, \ell_{\max}, |\hat{\mathcal{Y}}|)$ time. With the labels calculated, each of these calls has run time and sample complexity at most $\text{poly}(\log(|\mathcal{H}|, 1/\epsilon, \ell_{\max}, |\hat{\mathcal{Y}}|))$.

To audit for the DOI guarantee over the $(h, \ell) \in \{(f_{\ell,t}, \ell) : \ell \in \mathcal{L}\}$, at each iteration, we use the naive strategy outlined in Section 5.4 where we enumerate over all $|\mathcal{L}|$ losses and evaluate the performative risk of each pair $(f_{\ell,t}, \ell)$ on a single dataset of RCT samples of size $O(\ell_{\max}^2/\epsilon^2|\hat{\mathcal{Y}}|^2\log(|\mathcal{L}|))$ as per Corollary 5.4.4. Each auditing step for DOI therefore runs in time $|\mathcal{L}| \cdot \text{poly}(1/\epsilon, \ell_{\max}, |\hat{\mathcal{Y}}|)$ and uses $\text{poly}(1/\epsilon, \ell_{\max}, |\hat{\mathcal{Y}}|, \log(|\mathcal{L}|))$ many samples. All calls to the intermediate predictors $p^{(t)}$ also run in polynomial time as per Theorem 5.4.2. The final guarantees come from multiplying the sample and run time complexity of each iteration of the POI-boost algorithm by the bound on the total number of iterations. □

[9] Here, we have avoided discussion on the failure probability parameter δ for the $\mathcal{A}_{\text{csc}}$. However, it is clear that the relevant complexity bounds should depend only on $\log(1/\delta)$ and that applying a simple union bound would not change the nature of the resulting analysis. We therefore assume that the algorithms succeed with probability 1 for the sake of simplicity.

The main take away from this result is that if the cost-sensitive classification problem can be solved efficiently, in the sense that the the relevant statistical and computational complexities scale as polylog$|\mathcal{H}|$, then the overall POI-Boost algorithm runs in time linear in $|\mathcal{L}|$, poly-logarithmically in the size of $\mathcal{H}$ and with at most polylog$|\mathcal{H}||\mathcal{L}|$ many samples. Therefore, we can hope to develop efficient omnipredictors that are optimal for exponentially many decision rules, and polynomially many losses.

Note that, because of the result outlined in Proposition 5.3.3, this theorem also bounds the statistical and computational complexity of achieving universally adaptable omnipredictors. More specifically, the number of samples and the runtime for achieving universally adaptable omnipredictors are also bounded by the quantities in Corollary 5.4.8 where we now replace the class $\mathcal{L}$ by augment collection $\mathcal{L}_{\mathcal{W}}$ as defined in Proposition 5.3.3. The main difference is that the relevant runtime and sample complexity bounds replace dependence on $|\mathcal{L}|$ by dependence on $|\mathcal{L}||\mathcal{W}|$ and replace dependence on $\ell_{\max}$ by $\ell_{\max}\omega_{\max}$. Here, $\omega_{\max}$ is the the worst case density ratio for the class $\mathcal{W}$.

$$\max_{\omega\in\mathcal{W}} \max_{x\in\text{supp}(\mathcal{D}_\omega)} \omega(x) = \frac{\mathcal{D}_\omega(x)}{\mathcal{D}(x)}.$$

This complexity measure capture the intuition that if individuals x are poorly represented over the distribution $\mathcal{D}$ we are learning over, then we need more samples (and consequently runtime), to learn universally adaptable omnipredictors. We think of these complexity parameters like $\omega_{\max}$ as a first step. It is an interesting question for future work to provide sharper notions of problem complexity and to find ways of designing omnipredictors for exponentially large collections of importance weights $\mathcal{W}$.

5.5 Connections to Multicalibration

So far, we have studied how extensions of outcome indistinguishability definitions enable the design of omnipredictors for performative settings. In the world of supervised learning, [18] established tight connections between outcome indistinguishability and various notions of multicalibration [34]. Given the complementary relationship between these two concepts in the supervised world, it is natural to speculate that generalizing multicalibration to the outcome performative setting might be fruitful.

In this section, we begin to examine these questions and discuss analogues of multiaccuracy and multicalibration for the performative setting. We start by showing that multiaccuracy naturally, and efficiently, extends to performative contexts, and provides an effective way to achieve performative outcome indistinguishability in a loss-independent fashion. Conversely, we illustrate how naive translations of multicalibration to performative

prediction result in definitions whose complexity blows up exponentially in the number of predictions $|\hat{\mathcal{Y}}|$. We conclude with some discussion of alternatives to calibration-style guarantees that could, in principle, be used to obtain efficient omnipredictors.

On Multiaccuracy. Theorem 5.2.5 shows how $(\mathcal{L}, \mathcal{H}, \epsilon)$-performative omniprediction arises as a consequence of $(\mathcal{L}, \mathcal{H}, \epsilon)$-performative OI and $(\mathcal{L}, \epsilon)$-decision OI, where the OI distinguishers explicitly account for the collection of loss functions $\mathcal{L}$. Here, we show an efficient approach for obtaining POI for the class of all bounded input-oblivious loss functions.

We say that a loss function is *input-oblivious* if it only depends on the input $x \in \mathcal{X}$ via the decision $h(x)$. That is, for all x and x' and pairs $(\hat{y}, y)$, $\ell(x, \hat{y}, y) = \ell(x', \hat{y}, y)$. Equivalently, these functions have domain $\hat{\mathcal{Y}} \times \mathcal{Y}$ instead of $\mathcal{X} \times \hat{\mathcal{Y}} \times \mathcal{Y}$. We use

$$\mathcal{L}_{\text{io}} = \{\ell : \hat{\mathcal{Y}} \times \{0, 1\} \to [0, 1]\}$$

to denote the class of all bounded input-oblivious loss functions. Our first result for this section proves that a performative analogue of multiaccuracy [34, 41] implies POI for $\mathcal{L}_{\text{io}}$.

Definition 5.5.1 (Multiaccuracy)**.** For a distribution $\mathcal{D}$, hypothesis class $\mathcal{H}$, and $\epsilon \geq 0$, a predictor $\tilde{p} : \mathcal{X} \times \hat{\mathcal{Y}} \to [0, 1]$ is $(\mathcal{H}, \epsilon)$-multiaccurate under outcome performativity over $\mathcal{D}$ if for all $h \in \mathcal{H}$ and $\hat{y} \in \hat{\mathcal{Y}}$

$$\left|\mathbb{E}_{\substack{x \sim \mathcal{D} \\ y^* \sim p^*(x, \hat{y})}} [y^* \cdot \mathbf{1}\{h(x) = \hat{y}\}] - \mathbb{E}_{\substack{x \sim \mathcal{D} \\ \tilde{y} \sim \tilde{p}(x, \hat{y})}} [\tilde{y} \cdot \mathbf{1}\{h(x) = \hat{y}\}]\right| \leq \epsilon.$$

Here, we require that the expectation of our modeled outcome $\tilde{y} \sim \tilde{p}(x, h(x))$ is accurate after deploying each $h \in \mathcal{H}$, even when restricting our attention to the individuals $x \in \mathcal{X}$ such that $h(x) = \hat{y}$. While seemingly simpler than performative OI, we show that multiaccuracy in fact implies performative OI for all input-oblivious losses.

Lemma 5.5.2. *If $\tilde{p} : \mathcal{X} \times \hat{\mathcal{Y}} \to [0, 1]$ is $(\mathcal{H}, \epsilon)$-multiaccurate, then $\tilde{p}$ is $(\mathcal{L}_{\text{io}}, \mathcal{H}, 2\epsilon)$-performative OI.*

Proof. The proof shows an approximate equality between the loss $\ell \in \mathcal{L}_{\text{io}}$ of any $h \in \mathcal{H}$

under $\tilde{y} \sim \tilde{p}(x, h(x))$ and $y^* \sim p^*(x, h(x))$. For $\mathcal{Y} = \{0, 1\}$, $\mathbb{E}_{\substack{x\sim\mathcal{D} \\ \tilde{y}\sim\tilde{p}(x,\hat{y})}}[\ell(h(x), \tilde{y})]$ satisfies:

$$
\begin{aligned}
&= \sum_{\hat{y}\in\hat{\mathcal{Y}}} \Pr_{\mathcal{D}}[h(x) = \hat{y}] \cdot \mathbb{E}_{\substack{x\sim\mathcal{D} \\ \tilde{y}\sim\tilde{p}(x,\hat{y})}}[\ell(\hat{y}, \tilde{y}) \mid h(x) = \hat{y}] \\
&= \sum_{\hat{y}\in\hat{\mathcal{Y}}} \Pr_{\mathcal{D}}[h(x) = \hat{y}] \cdot \mathbb{E}_{\substack{x\sim\mathcal{D} \\ \tilde{y}\sim\tilde{p}(x,\hat{y})}}[\tilde{y} \cdot \ell(\hat{y}, 1) + (1 - \tilde{y}) \cdot \ell(\hat{y}, 0) \mid h(x) = \hat{y}] \\
&\leq \sum_{\hat{y}\in\hat{\mathcal{Y}}} \Pr_{\mathcal{D}}[h(x) = \hat{y}] \cdot \mathbb{E}_{\substack{x\sim\mathcal{D} \\ y^*\sim p^*(x,\hat{y})}}[y^* \cdot \ell(\hat{y}, 1) + (1 - y^*) \cdot \ell(\hat{y}, 0) \mid h(x) = \hat{y}] + 2\epsilon \\
&= \sum_{\hat{y}\in\hat{\mathcal{Y}}} \Pr_{\mathcal{D}}[h(x) = \hat{y}] \cdot \mathbb{E}_{\substack{x\sim\mathcal{D} \\ y^*\sim p^*(x,\hat{y})}}[\ell(\hat{y}, y^*) \mid h(x) = \hat{y}] + 2\epsilon \\
&= \mathbb{E}_{\substack{x\sim\mathcal{D} \\ y^*\sim p^*(x,\hat{y})}}[\ell(h(x), y^*)] + 2\epsilon.
\end{aligned}
$$

The the third line follows under the assumption that $\tilde{p}$ is $(\mathcal{H}, \epsilon)$-multiaccurate and the bound on the magnitude of $|\ell(\hat{y}, b)| \leq 1$ for all $\hat{y} \in \hat{\mathcal{Y}}$ and $b \in \{0, 1\}$. Given that an identical argument can be used to show the opposite inequality, we conclude that $\tilde{p}$ is indeed POI. □

Inspecting the performative multiaccuracy condition, we can see that it is similarly possible to reduce the problem of auditing for multiaccuracy to supervised learning. This auditing procedure can be viewed as a special case of the auditing step from Section 5.4 or as a generalization of previous auditing procedures from work on multiaccuracy in the supervised learning setting [34, 41]. In more detail, the relevant auditing problem for performative multiaccuracy is to determine whether there exists an $h \in \mathcal{H}$ and $\hat{y} \in \hat{\mathcal{Y}}$ such that

$$
\left|\mathbb{E}_{\substack{x\sim\mathcal{D} \\ y^*\sim p^*(x,\hat{y})}}[(y^* - \tilde{p}(x, \hat{y})) \cdot \mathbf{1}\{h(x) = \hat{y}\}]\right| > \epsilon.
$$

As before, this auditing step reduces to a cost-sensitive classification problem (Definition 5.4.6). Auditing over a hypothesis class $\mathcal{H}$ can be done with $2|\hat{\mathcal{Y}}|$ many calls to a cost-sensitive learner $\mathcal{A}_{\text{csc}}$ with tolerance parameter $O(\epsilon)$. We omit a formal statement of this result since it follows the exact pattern from Proposition 5.4.7.

In other words, in order to achieve $(\mathcal{L}, \mathcal{H}, O(\epsilon))$-POI for any class of input-oblivious losses $\mathcal{L} \subseteq \mathcal{L}_{\text{io}}$, it suffices to audit for and enforce $(\mathcal{H}, \epsilon)$-multiaccuracy. In this sense, for input-oblivious losses, there is a single auditing procedure that works for all losses, so we can replace $|\mathcal{L}|$-factors by $|\hat{\mathcal{Y}}|$ factors in the auditing complexity for performative OI.

On Multicalibration. Going beyond multiaccuracy, the original work of [29] established omniprediction in the supervised setting as a consequence of multicalibration. As such, we might wonder whether there exists an analogous notion of multicalibration for performative prediction that enables a similar result. Defining an efficient notion of calibration under performativity (let alone, multicalibration), turns out to be a subtle task.

In supervised learning, calibration requires that the expectation of $\tilde{p}$ is accurate, even when we partition the inputs $x \in \mathcal{X}$ based on the predicted value $\tilde{p}(x) = v$. Specifically, the constraints quantify over each supported $v \in \text{supp}(\tilde{p})$:

$$\mathbb{E}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x)}}[y^* \cdot \mathbf{1}\{\tilde{p}(x) = v\}] \approx \mathbb{E}_{\substack{x\sim\mathcal{D}\\ \tilde{y}\sim\tilde{p}(x)}}[\tilde{y} \cdot \mathbf{1}\{\tilde{p}(x) = v\}] = v \cdot \Pr[\tilde{p}(x) = v].$$

Such calibration-style constraints suffice to establish omniprediction because for any loss ℓ, the optimal decision $\tilde{f}_\ell(x)$ is completely determined by $\tilde{p}(x)$.

In performative prediction, quantifying over the supported values of $\tilde{p}$ requires considering the decisions $\hat{y} \in \hat{\mathcal{Y}}$ as well. In particular, the optimal decision $\tilde{f}_\ell(x)$ is a function of the vector of predictions $\tilde{q}(x) \in [0,1]^{|\hat{\mathcal{Y}}|}$, which gives the predicted probability $\tilde{y} \sim \tilde{p}(x, \hat{y})$ for each $\hat{y} \in \hat{\mathcal{Y}}$ (recall the $q(\cdot)$ notation from Proposition 5.4.1). Thus, using the naive translation of the calibration constraints for omniprediction, we must partition $\mathcal{X}$ based on the vector-valued predictions, $\tilde{q}(x) = \overrightarrow{v}$. The cardinality of this calibration partition of $\mathcal{X}$ scales exponentially in the number of decisions $|\hat{\mathcal{Y}}|$, even for binary outcomes $\mathcal{Y}$.

Still, we may consider more efficient calibration-style conditions that suffice to imply omniprediction in the performative setting. Rather than aiming for full performative calibration, we focus on adapting the notion of decision calibration [89] to the performative setting. Decision calibration was introduced to avoid exponential blow-up in the calibration constraints due to multi-class prediction. We show that the notion can equally be adapted to the performative setting to deal with blow-up due to many actions $\hat{y} \in \hat{\mathcal{Y}}$. We define decision calibration with respect to the class of input-oblivious losses.

Definition 5.5.3 (Decision Calibration)**.** For a distribution $\mathcal{D}$ and $\epsilon \geq 0$, a predictor $\tilde{p} : \mathcal{X} \times \hat{\mathcal{Y}} \to [0,1]$ is ϵ-decision calibrated under outcome performativity over $\mathcal{D}$ if for every loss $\ell \in \mathcal{L}_{\text{io}}$, and for all $\hat{y} \in \hat{\mathcal{Y}}$,

$$\left|\mathbb{E}_{\substack{x\sim\mathcal{D}\\ y^*\sim p^*(x,\tilde{f}_\ell(x))}}[y^* \cdot \mathbf{1}\{\tilde{f}_\ell(x) = \hat{y}\}] - \mathbb{E}_{\substack{x\sim\mathcal{D}\\ \tilde{y}\sim\tilde{p}(x,\tilde{f}_\ell(x))}}[\tilde{y} \cdot \mathbf{1}\{\tilde{f}_\ell(x) = \hat{y}\}]\right| \leq \epsilon.$$

With this definition in place, the proof of Lemma 5.5.2 can be adapted to show that decision calibration suffices to establish performative decision OI.

Lemma 5.5.4. *If $\tilde{p} : \mathcal{X} \times \hat{\mathcal{Y}} \to [0,1]$ is ϵ-performative decision calibrated, then $\tilde{p}$ is $(\mathcal{L}_{\text{io}}, 2\epsilon)$-performative decision OI.*

As an immediate corollary of Theorem 5.2.5 and Lemmas 5.5.2 & 5.5.4, we obtain sufficient conditions for omniprediction with respect to all bounded input-oblivious losses.

Corollary 5.5.5. *Suppose $\tilde{p} : \mathcal{X} \times \hat{\mathcal{Y}} \to [0,1]$ is $(\mathcal{H}, \epsilon)$-multiaccurate and ϵ-decision calibrated under outcome performativity. Then, $\tilde{p}$ is a $(\mathcal{L}_{\text{io}}, \mathcal{H}, 4\epsilon)$-performative omnipredictor.*

In other words, if we can obtain multiaccuracy and decision calibration under performativity, then we have a direct pathway to obtain omniprediction for all bounded, input-oblivious losses.

On Decision Calibration. Note, however, that unlike the case of multiaccuracy, the decision rules that arise in the decision calibration condition are loss-dependent. That is, the optimal decision rules $\tilde{f}_\ell$ depend on $\ell \in \mathcal{L}_{\text{io}}$.

Motivated by the strong guarantee, we consider the feasibility of auditing for decision calibration. Note that for any $\ell \in \mathcal{L}_{\text{io}}$, the loss is defined by the loss for each outcome $y \in \{0,1\}$ and decision $\hat{y} \in \hat{\mathcal{Y}}$. Thus, to audit decision calibration over all input-oblivious losses, it suffices to audit whether there exist a $\hat{y} \in \hat{\mathcal{Y}}$ and $w_{a,0}, w_{a,1} \in [-1,1]$ such that:

$$\left|\mathbb{E}_x[(\tilde{p}(x,\hat{y}) - y^*) \cdot \mathbf{1}\{\arg\min_{a \in \hat{\mathcal{Y}}}\{w_{a,1} \cdot \tilde{p}(x,a) + w_{a,0} \cdot (1 - \tilde{p}(x,a))\} = \hat{y}\}]\right| > \epsilon$$

where the weights $w_{a,0}$ and $w_{a,1} \in [-1,1]$ represent the choice of $\ell(a,0)$ and $\ell(a,1)$ corresponding to the choice of $a \in \hat{\mathcal{Y}}$.

Naively, searching for such a violated loss might require time exponential in $|\hat{\mathcal{Y}}|$. For instance, by explicitly enumerating over some appropriately-fine net of $[-1,1]^{2|\hat{\mathcal{Y}}|}$, then we can simply consider "every" possible loss. Improving the computational complexity of such a search is an interesting question, which may benefit from the techniques utilized in [89].

Even without an improvement in the runtime complexity of learning, note that once an auditor succeeds, and we have a violated loss function, there is an efficient update to the prediction function. In particular, we simply need to record the chosen $2 \cdot |\hat{\mathcal{Y}}|$ parameters $\{w_{a,0}, w_{a,1}\}$ and the $\hat{y} \in \hat{\mathcal{Y}}$, then execute the update from POI-Boost. In all, we can conclude that performative omnipredictors for $\mathcal{L}_{\text{io}}$ exist in complexity independent of the complexity of $\mathcal{L}_{\text{io}}$.

Corollary 5.5.6. *Suppose $\mathcal{H}$ is a hypothesis class with size-s circuits. Then, for any $\epsilon > 0$, there exist an $(\mathcal{L}_{\text{io}}, \mathcal{H}, \epsilon)$-performative omnipredictor implemented by a circuit of size* $\text{poly}(s, |\hat{\mathcal{Y}}|)/\epsilon^2$.

This preliminary analysis leaves open the possibility of learning performative omnipredictors via techniques that are independent of the loss class, as in the original work

on $(\mathcal{L}_{\text{cvx}}, \mathcal{H})$-omniprediction from $\mathcal{H}$-multicalibration. We leave a more thorough investigation of these ideas to future work.

5.6 Chapter Notes

The material from this chapter is derived from [42]. The initial definition of omniprediction was introduced in [29], in which the authors proved how it could be achieved as a consequence of multi-calibration [34]. Later work established a connection between omniprediction and outcome indistinguishability [28], a perspective we extend in this chapter.

Part II

Empirical Investigations

Foreword to Part II

In the second part of this thesis, we shift our focus from developing the theoretical foundations of performative prediction to understanding how social prediction dynamics play out empirically. On a personal note, the research presented in this next chapter was initially motivated by our prior belief that educational predictions were strongly performative and that we should be studying these prediction problems as such.

Needless to say, we were quite surprised to see that early warning systems in Wisconsin are not performative at all, hence the title of the chapter. Due to an accumulation of social and structural factors present in public schools, there is a large gap between developing insightful predictions, and translating those predictions into interventions that meaningfully change the life trajectories of students in schools. We will exactly delve into exactly why this is over the remainder of the thesis.

However, before getting into the results of this investigation, we pause to discuss how the two parts of this thesis connect and our discuss views on the relationship between theory and practice. Perhaps the biggest impact of our earlier theoretical work on these empirical investigations is that it has provided a guiding lens to hopefully ask the right kinds of questions. Mathematics on some level is nothing but the art of formal reasoning. Although it may not always be transparent to see, the earlier theoretical investigations on retraining and the comparisons between stability and optimality greatly helped inform our views on prediction in public high schools and what were the things that we should look out for.

Furthermore, while this particular example of social prediction may not have been performative and some theorems may not apply, we believe that it is the job of the theorist to be forward-looking. As computation and data become more a core part of our daily lives, there will be more examples of feedback loops between prediction systems and their encompassing environments. We sincerely hope that our efforts in developing the theoretical foundations of performative prediction will guide future empirical research into other problem domains where predictions do indeed have strong consequences on the observed data.

Lastly, the connections between theory and practice go both ways. Not only has the theory helped guide empirical inquiry, the empirical results have already started inspired further theoretical research. While we weren't able to include these latest results into the thesis, we find that this collaboration with the Wisconsin DPI has been inspiring and helped ground theoretical questions in concrete empirical applications.

Chapter 6

Difficult Lessons on Social Prediction from Wisconsin Public Schools

6.1 Introduction

A class of automated risk prediction tools known as early warning systems (EWS) has recently become part of the de facto approach towards tackling low high school graduation rates across the United States. EWS were built in part as a response to the so-called "dropout crisis" of the early 2000s and are fueled by data collection efforts resulting from the 2001 No Child Left Behind Act. These tools typically use data about students, schools, and districts to predict each student's dropout risk. After initial programs in Chicago and Philadelphia during the late 2000s, EWS boomed in use across the nation [4, 8]. By 2015, over half of US public schools had implemented some version of an EWS, according to a survey by the Department of Education [82].

Early warning systems aim to identify potential high-risk students early to assist educators in effectively targeting interventions to individual students [49]. Despite their surge in popularity and significant financial investment by education departments across the country, we lack conclusive evidence regarding the efficacy of predictive systems in reducing student dropout. Challenges in setting up such empirical studies and time lags involved in measuring the impact of early interventions on later high school graduation rates pose key barriers. As a result, existing studies on EWS focus on short-term impacts, are conducted on relatively small sample sizes, and present inconclusive results on the bottom-line effect on graduation rates [7].

In this work, we evaluate the decade-long impact of an early warning system—Dropout Early Warning System (DEWS)—designed by the Wisconsin Department of Public Instruction (DPI) and used throughout the state's public schools. Using an order of magnitude

more data than previous studies, we show that the system's predictions are highly accurate assessments of the true probability with which individual students will drop out of high school. These findings hold even when limiting our analyses to students from marginalized groups. Nonetheless, we find no evidence that DEWS has improved graduation outcomes, even when we restrict ourselves to schools that are active users of the system.

These two insights—high predictive accuracy on the one hand and non-effect on graduation outcomes on the other—beg for an explanation. A reasonable conjecture is that schools lack sufficient instructions on how to translate risk predictions into effective interventions. DPI has certainly expended effort to increase the use of DEWS, as well as to identify and implement educational interventions. Amongst other initiatives, DPI has collaborated with organizations such as the Wisconsin Response to Intervention Center to organize on-site DEWS training for districts throughout the state, as well as maintaining an expansive set of online resources. However, there remains room for improvement in providing comprehensive instructions about using DEWS as an intervention tool. This conjecture is plausibly part of the picture. Yet it also calls for a deeper study of educators in situ.

We show that a stronger force is at play that explains the observations we see. Specifically, we identify a robust statistical pattern that we summarize as:

Academic outcomes are essentially statistically independent of individuals conditional on their environments.

Here, individual features are measurements that directly correspond to a specific student. Examples include student test scores, socioeconomic status, misconduct records, and days of absence. Environmental features, on the other hand, describe schools and districts rather than a single student. Examples include school size, financial budget, and districts' aggregate socioeconomic and demographic statistics. DEWS uses both individual and environmental features to make predictions about individual students.

This statistical independence suggests an explanation of our observations about DEWS. Because individual graduation outcomes are strongly correlated with the available set of features (and in particular, the environmental features), a model that uses these statistics to predict individual dropout is highly accurate when evaluated across the *entire* population of students in Wisconsin public schools. However, conditioning on students belonging to the *same* school, outcomes are largely independent of the available individual features. Hence, predictions are not much better within each school than randomly guessing which students are at risk of dropping out. Even if a particular school has effective educational interventions, administering them on the basis of DEWS scores is no better than a random allocation of the intervention.

Summary of our Methods and Analyses

We first do a deep-dive into the prediction model to establish that the risk assessments made by the system are highly accurate. Over the past decade, nearly 97% of students identified as low-risk graduated from high school on time, while only 70% of students in the high-risk group completed their degrees within four years. These gaps are significant relative to the 90% high school completion rate for the state overall. This is in contrast to work highlighting challenges in predicting such life outcomes [31, 73, 74]. Further, contrary to several previous studies on prediction systems in social settings [**obermeyer**, 21, 27], DEWS scores are a *more* accurate assessment of the true dropout risk for students from under-served and marginalized backgrounds (e.g., Black and Latino students).

We find, however, that these accurate predictions do not translate to effective interventions that increase on-time graduation. We establish this observation via a regression discontinuity design: We estimate that assigning students into higher predicted risk categories improves their chances of graduation by less than 5%. Nevertheless, the 95% confidence interval for this estimated effect firmly includes zero.

A primary concern resulting from these insights is whether and how schools use the system. After its roll-out in 2012, schools quickly adopted DEWS. Tracking visits to the online DEWS portal, we find that around two-third of schools regularly log onto the platform. Furthermore, usage is concentrated amongst the more populated school districts with below-average graduation rates and a higher percentage of students from marginalized backgrounds. To the extent that we can test this with available data, usage does not appear to provide an explanation for the above findings.

Our proposed empirical explanation for the non-effect results comes from examining the mechanisms that make dropout predictable in the first place. In our analyses, we discover a robust statistical pattern pervasive throughout Wisconsin public schools. Namely, environmental features that are defined at the level of schools and districts, contain significant signal about dropout risk across the population. However, *within* the same school environment, graduation outcomes are almost entirely unrelated to the extensive set of individual student features, including race, gender, and test scores.

This empirical observation serves as a compelling explanation in part because it can be tested and possibly refuted. First, if outcomes are approximately independent of individual features, including these individual features in a predictive model that already uses environmental information should only lead to a negligible improvement in accuracy. We test this implication empirically and find that adding individual features to an environment-based predictor improves mean squared error by around 10%. These improvements are relatively minor within the student population at high risk of leaving high school prior to earning their diplomas.

Furthermore, the independence of individuals and graduation outcomes conditioned on the environment guarantees that schools must be statistically homogeneous groups of students with respect to their dropout risk. We test this second implication and find that the predictor that uses individual features assigns all students within the same school a near-identical probability of graduation. Our results show this is essentially the optimal prediction given the available data, since schools have minimal variation in individual features. For example, most variation in state-wide standardized exam scores comes from students in different schools scoring very differently. However, students in the same school tend to have similar test scores.

Implications

The fundamental empirical fact we discover takes precedence over questions about the effectiveness and availability of educational interventions as the primary concern. Indeed, DEWS, as a tool aimed at reducing dropout by targeting individual students, would continue to be ineffective regardless of the strength of educational interventions available to schools. To see this suppose, as a thought-experiment, that all schools have an intervention that perfectly mitigates student dropout in any student to whom it is applied. Suppose that the intervention has a fixed cost and that schools have a fixed budget to spend. Finally, suppose that schools allocate this hypothetical intervention at least as well as a random allocation. Under these assumptions, our statistical finding implies that DEWS will not make any improvement over the school's current allocation of the intervention. The reason is that, conditional on school, DEWS predictions are independent of dropout. As such, the allocation of the intervention suggested by DEWS is is no better than random.

Our findings have immediate implications around the design of early warning systems and viability of interventions assigned on the basis of statistical risk scores. First, they challenge the extent to which dropout is an unpredictable life outcome versus a highly foreseeable event. Contrary to the commonly-held belief that important life outcomes are inherently unpredictable [73], the predictive strength of environmental, time-invariant features proves that future graduation outcomes can be reliably forecast from the moment children enroll in kindergarten. This also suggests that additional resources can be effectively targeted towards marginalized students at a young age.

Second, our findings dispel the belief that predictability implies improvement. Even though DEWS scores are highly accurate forecasts of the true individual probability of on-time graduation, there is no evidence that this predictability of future outcomes has translated into more students in Wisconsin finishing their high school degrees.

Lastly, these results illustrate how sophisticated statistical algorithms may be of limited value in settings where risk assessments are made at the level of individuals but outcomes

are significantly driven by structural forces. In our case, the overarching goal of an early warning system is to serve as an efficient targeting mechanism. An early warning system should precisely answer the question: *"which students need the most help?"* One of our central findings is that sophisticated prediction algorithms provide little additional insight into this question beyond simply ranking students according to the average graduation rate in their school district.

Combined, our findings provide a novel statistical and empirical backbone for a robust qualitative insight from the education policy and advocacy community: The bottleneck is not how to identify students at high risk of dropping out, but rather how to overcome structural barriers to accessing well-resourced schools and neighborhoods. In the case of Wisconsin, dropout is disproportionately concentrated in working-class, urban districts where the overwhelming majority of students are Black or Latino. In fact, five schools account for 10% of all dropouts in the state. Bringing graduation rates within these schools to the state average would have an outsize impact on the dropout problem. Compared to investing in other community-based social service interventions, the decision to fund and implement sophisticated early warning systems without also devoting resources to interventions tackling structural barriers should be carefully evaluated in light of these school-level disparities.

6.2 Background

Early Warning Systems - History & Previous Research

Early warning systems emerged as a critical public response to the so-called "dropout crisis" of the early 2000s. During this period, there was widespread, bipartisan recognition of alarmingly high numbers of young adults without high school degrees, especially amongst under-served populations [83]. These concerns led Congress to pass the No Child Left Behind Act of 2001. NCLB is credited with increased data collection efforts, often tied to school accountability.

Following this legislation and increased data availability, EWS boomed after the publication of two pilot program studies in 2007 [4, 8]. These papers demonstrated the potential for the existence of high-fidelity dropout indicators amongst high-poverty, majority Black and Latino public schools in Chicago and Philadelphia. A primary outcome of these studies was the creation of the so-called "ABC" indicators of dropout (attendance, behavior, & coursework), which the authors argue are both simple to measure and can accurately identify future dropouts [12]. Following these pilot programs, EWS quickly became mainstream: 52% of respondents to a survey organized by the U.S Federal De-

partment of Education said that they had implemented some version of an EWS by the 2014-15 academic year [82].

Over the last decade, there have been various observational studies and randomized control trials evaluating the effectiveness of EWS. In 2014, the U.S. Department of Education conducted a randomized control trial with 73 schools containing roughly 38,000 students from three Midwestern states [24]. Approximately half of the schools were randomly assigned to implement an EWS. After a year of "limited implementation," the study found that schools in the treatment group had a small reduction in chronic absenteeism. However, they observed no effects across other measured outcomes such as number of students suspended or the fraction of students with relatively low grade point averages. [49] conducted a similar randomized control trial across 41 schools over two years and reached similar conclusions.

Despite the lack of strong empirical backing, some experts remain consistently optimistic about the efficacy of EWS, and several states have continued to invest heavily in their development (e.g., Massachusetts). Summarizing a consensus in the field, [7] states that given the short time horizons and small sample sizes, "overall, evidence gathering is still in the early stages, promising but not fully confirmed." Using state-wide data on over 220,000 students from 2013 until 2021, our work addresses this gap in the education literature as the first large-scale evaluation of the long-term relationship between the use of EWS and the likelihood of on-time high school graduation.

Wisconsin Public Schools and the DEWS Program

Wisconsin has one of the highest high school graduation rates in the United States, with around 90% of its students receiving their degrees within the expected four years. Despite this high graduation rate, the state also has one of the largest gaps between demographic groups. For instance, in the 2019-2020 academic year, over 94% of White students graduated from high school on time, but less than 75% of Black students completed high school within four years, in part due to disparities in the quality of education available to Black and White students.[1] Addressing these educational disparities has been a significant focus of the department. There are a total of around 65,000 students per grade in Wisconsin public schools, with most of the Black and Latino populations concentrated within the larger, urban districts.[2]

Launched in 2012 by the Department of Public Instruction, the Dropout Early Warning System (DEWS) estimates the likelihood that each public middle school student (grades 5 through 8) in Wisconsin will graduate high school on time. Scores are generated at

[1] See here for a complete breakdown of graduation rates by state and demographic groups.

[2] More information on school demographics may be found on the DPI website.

the beginning of each academic year by the DPI and published to an online platform (WISEDash) where administrators can voluntarily log on to see the DEWS predictions for all students in their district. As discussed previously, website visit statistics provided by the DPI show that over two-thirds of districts regularly use the DEWS system. Furthermore, usage is higher amongst lower-income districts with higher percentages of Black and Latino students. These districts also tend to have below-average graduation rates. For further details on system usage, please see the supplementary material for this chapter.

Predictions are primarily used by school counselors in Wisconsin public high schools as a means of triaging new student cohorts as they enter 9th grade (personal comm). The department recommends, but does not mandate, that schools focus their resources on students assigned to the high-risk category and then move on to evaluating lower-risk students. Please see the DEWS action guide for a more comprehensive description of the DPI recommendations regarding how to use and interpret DEWS predictions.

On an implementation level, the system generates predictions using over 40 student features. These encompass a wide range of areas, including demographic and socioeconomic information (e.g., race, gender, family income), academic performance (e.g., scores on state standardized exams), as well as community-level statistics (e.g., percent of cohort that is non-White and school size).[3]

The main outputs of the EWS are 1) the DEWS risk category (or label) and 2) the DEWS score. The DEWS category takes one of three values: low, moderate, or high. It provides a simple way to interpret a holistic assessment of dropout risk for individual students. On the other hand, the DEWS score is an estimated probability of on-time graduation. It takes continuous values between 0 and 1.

New models are trained every academic year for every grade. The models are ensembles of state-of-the-art supervised learning methods for tabular data, such as random forests and gradient-boosted decision trees. Models are fit via empirical risk minimization on a dataset of a historical cohort of students. These datasets consist of features and outcomes for the five most recent cohorts of students in a specific grade and for whom graduation outcomes have been observed. For example, the 8th-grade model for the 2020-2021 academic year is trained using data from the 8th-grade cohorts between the 2011-2012 and 2015-16 years.

For further background on the DEWS system, please see the initial white paper on the program [44], or visit the Wisconsin DPI website.

[3]Please see the supplementary material for a more comprehensive description of the features used by the system.

Evaluation Framework

From an algorithms point of view, it is helpful to think of an EWS as a sorting procedure. The system takes as input the population of students in the public school system and outputs a sorted list (i.e., ranking) of these students according to their need for effective educational interventions. In the case of the DEWS program, the probability of on-time graduation, or DEWS score, serves as the concrete proxy for this subjective notion of "need".

Given this ordering, the second key component of an EWS is to choose a threshold. The system prioritizes students whose probability of graduation is below this threshold to receive effective interventions, while students above this threshold are relatively deprioritized. In the DEWS program, this thresholding is implemented by the DEWS label. As we will explain later, thresholding the continuous DEWS score determines the DEWS label (the low, moderate, and high risk categories). According to the DEWS action guide, the DPI recommends that schools take immediate action on students who receive a high risk prediction. In contrast, if capacity allows, students who receive a moderate risk prediction should only be examined *after* the high risk students. Approximately 10% of students each year across the entire state are labeled to be at high risk of dropping out, while only 6% are predicted to be at moderate risk. The rest receive low risk predictions.

In light of these design choices, an effective EWS should have two main criteria for success. First, it should accurately rank students according to their intervention needs. Students with a lower probability of on-time graduation should be ranked before students with relatively higher probabilities. Furthermore, this ranking should be accurate not just overall but also for students from under-served and marginalized backgrounds. Second, the thresholding, and subsequent assignment into different risk categories, should lead to meaningful changes in the likelihood of graduation for individual students. In other words, DEWS predictions should be strongly performative. Students whose scores lie below the threshold are explicitly prioritized for additional attention by the school. Ideally, high-risk predictions should therefore be self-negating prophecies. Bumping up the risk prediction from moderate to high risk should *increase* the likelihood of on-time graduation. We organize our analysis of the DEWS system according to this evaluation framework.

6.3 Does DEWS Accurately Identify Dropout Risk?

The hypothesized success of early warning systems is predicated on accurate identification of future students at risk of leaving high school without a diploma. Consequently, we begin our evaluation by understanding the extent to which DEWS predictions meaningfully distinguish high and low risk students. As we noted previously, predictions are primarily

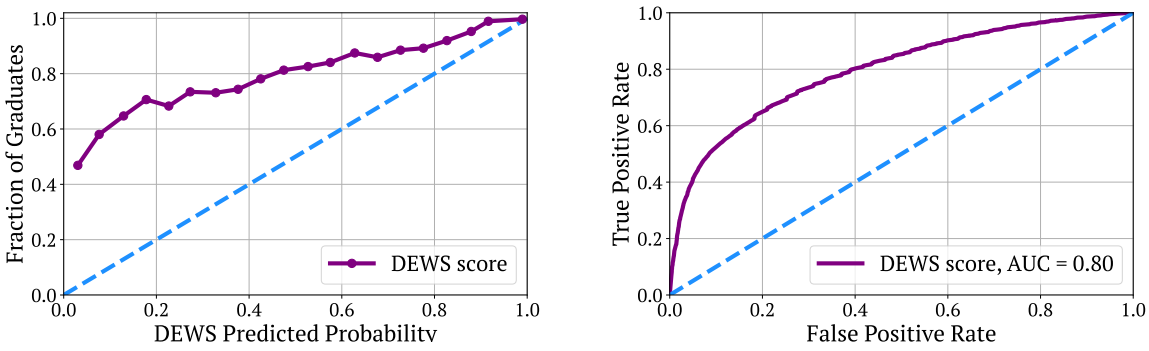


Figure 6.1: **Left:** calibration curve for the 8th grade DEWS scores (predicted probabilities of on-time graduation) from 2018-2021. **Right:** ROC curve for the same set of predictions and outcomes. We see that the scores are a highly accurate assessment of the relative risks of dropout for students in Wisconsin.

examined as students enter high school. Hence, we focus our analysis on the performance of the 8th grade DEWS model.

Starting from the 2013-14 academic year, the first year for which DEWS scores are available, until the last in 2021, there are 4 cohorts of 8th graders (roughly 215k students) for whom predictions were generated in 8th grade and for whom graduation outcomes were also observed. We focus our discussion on the extent to which DEWS has been predictive for this population.

First, we find that there are stark differences in graduation rates amongst students who were assigned into different risk categories. Nearly 97% of the students who were predicted to be low risk finished high school on-time, while less than 70% of students in the high-risk group graduated on-time. On the other hand, moderate risk students had an 83% graduation rate. For context, the overall graduation rate in Wisconsin is around 90%. From these gaps, we see that DEWS labels non-trivially categorize students into qualitatively meaningful risk groups.

Examining the continuous DEWS score provides a complementary perspective. In Figure 6.1, we present the calibration curve for these scores. Intuitively, calibration measures whether predictions are accurate. The plot illustrates, for students who were predicted to graduate on-time with a particular probability (measured on the x-axis), what fraction of those students actually graduated (y-axis). To give an example, a predictor is calibrated if amongst the students who were predicted to graduate with probability .9, roughly 90% graduate. Graphically, a predictor is perfectly calibrated if its calibration curve lies on the

diagonal $y = x$ line.

DEWS scores are as a whole miscalibrated; they consistently understate the true probability of on-time graduation. However, this miscalibration is mild in the sense that the DEWS scores are rank preserving. Students with higher predicted probabilities have higher graduation rates than those with lower predicted probabilities (the calibration curve is a line with positive slope). Hence, while the continuous scores are a misleading measure of *absolute* risk, they do provide an adequate assessment of *relative* risk. Again thinking of DEWS as a sorting algorithm, relative risk is in many ways a more relevant metric in education. The typical workflow for DEWS is that schools first rank students according to their scores and, in theory, focus their resources on students who are predicted to be at highest risk. Having an accurate measurement of relative risk ensures that students who need more help are more likely to receive attention.

Apart from their calibration curves, the predictive value of the continuous DEWS score is also evident from their induced receiver operating characteristic (ROC curve). These curves describe the entire set of possible true and false positive rates achievable by thresholding the continuous DEWS score. From the plot in Figure 6.1, we see that there exists a threshold such that predicting graduation outcomes from the DEWS score would identify nearly two thirds of all dropouts, while maintaining a false positive rate of less than 20%. The scores achieve a historical AUC (area under the curve) of .8. While lower than the initial AUC estimate of .86 predicted by earlier work on DEWS [44], these statistics further illustrate how DEWS scores are a non-trivial predictor of dropout.

Disparities in Predictive Accuracy

So far, we have established that DEWS scores are on average predictive when evaluated over the entire population and output an *overall*, accurate ranking of students according to their relative risks of dropping out. However, one of the stated goals of the system is not just to improve outcomes overall, but particularly amongst students from historically underserved groups [44].

To enable these improvements, it is crucial that predictions remain accurate when evaluation is restricted to students from marginalized demographic groups. Large disparities in predictive accuracy can, in principle, lead to unfair allocations of scarce school resources for mitigating dropout, thereby further exacerbating existing inequalities.

Our investigation into these concerns reveals mixed results. On the one hand, from the plots in Figure 6.2, we see that predictions for various historically underserved groups have lower calibration error than predictions for students in the majority. That is, the calibration curve for non-White, or low income students is relatively closer to the main $y = x$ diagonal. Therefore, DEWS scores provide a *more* accurate assessment of the absolute risk of dropout

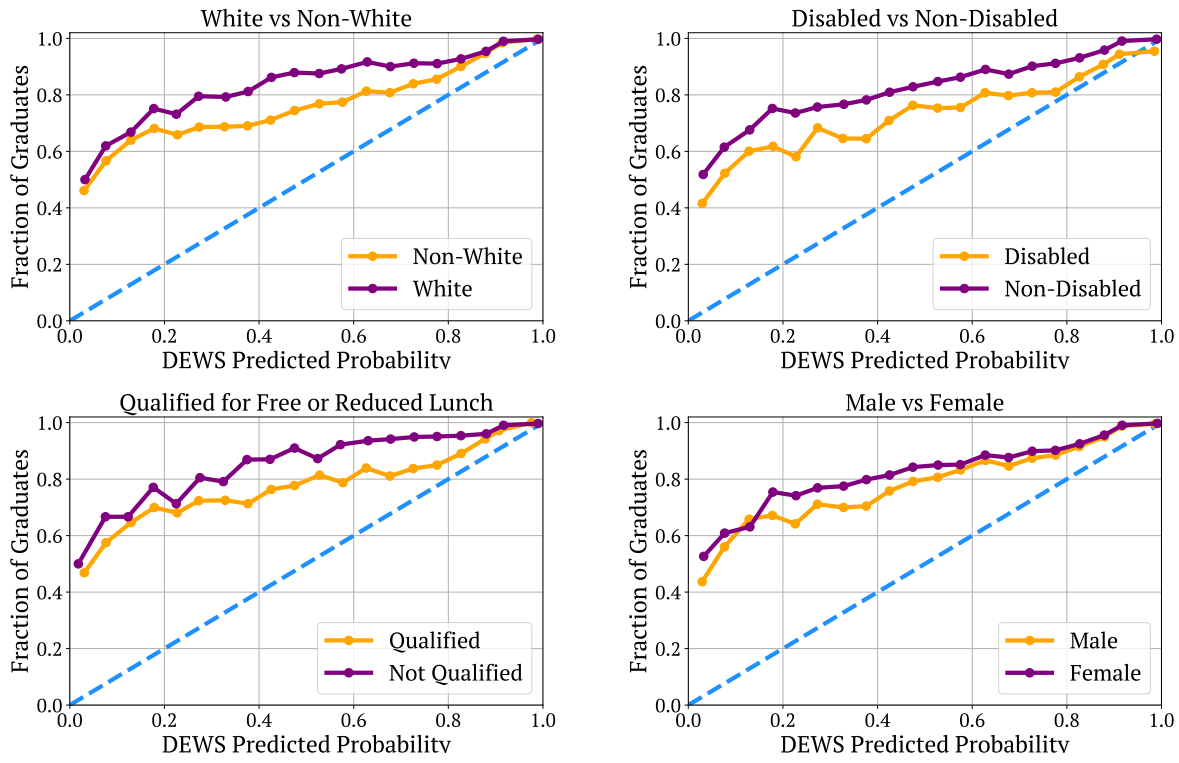


Figure 6.2: Comparing calibration curves for DEWS scores amongst different student groups. **Top Left:** comparison between white versus students of color. **Top Right:** comparison between students who have been diagnosed with a disability versus those that have not. **Bottom Left:** comparison based on qualifying for free or reduced lunch. **Bottom Right:** calibration comparison between male and female students. Overall, students from underserved groups tend to have lower calibration error.

for underserved students. On the other hand, from the perspective of relative risk, if schools select students for intervention by ranking their scores and selecting those with the lowest predicted probability of graduation, underserved students would be systematically overlooked and de-prioritized.

To see this, consider for example the top left plot comparing predictions on White versus non-White.[4] Amongst the population of students whose true rate of on-time

[4]The precise comparison here is between the group of students in the dataset that have a 1 in the column indicating whether a student is White and the group for which this feature is equal to 0. As per DPI practices, if this feature is 0, a student could be Black, Hispanic, Asian, Pacific Islander, Native American, or belong to two or more races. For the sake of being precise, we use the term non-White throughout the manuscript to

graduation was .8 (i.e y value equals .8), White students were assigned a DEWS score of around .35 while students of color had an assigned DEWS score closer to .6. Recall that DEWS scores are predicted probabilities of on-time graduation: lower scores indicate higher predicted risk. While both groups of students are equally at risk of dropping out (they share the same y value), if interventions are assigned on the basis of ranked DEWS scores, White students would be systematically intervened on before students of color. Furthermore, because the calibration curves for non-White students lie consistently below the curves for students in the majority, this behavior holds across the entire risk spectrum, not just for this 80% group.

As we will now see, DEWS scores have not had a significant impact on student graduation outcomes. Therefore, we find it highly unlikely that the program has exacerbated existing inequalities in academic achievement between demographic groups. However, schools that do choose to use an EWS going forward should be mindful of these predictive disparities when training their own predictors. We note that these can be easily fixed using recent advances from the algorithmic fairness literature [34, 66].

6.4 Do DEWS Predictions Lead to Better Graduation Outcomes? (Alternatively, Is DEWS Performative?)

So far, we have concluded that DEWS provides precise estimates of the future risk with which individual students will leave high school without a degree. Next, we examine whether this accurate sorting of students, and later thresholding into discrete risk categories, has translated into improved graduation outcomes.

Recall that predictions are generated with the very explicit *intent* of improving student outcomes. As discussed previously, the ideal early warning system should therefore have the property that high-risk predictions are self-negating. Students prioritized by the system to receive additional attention should graduate at higher rates than in a counterfactual world where they are predicted to be at low or moderate risk of dropping out.

Through a number of different analyses, we find no evidence that this is the case. Examining these highly accurate predictions has not led to significant improvements in the rate of on-time graduation for students in Wisconsin public schools.

We measure the causal impact of assigning students into higher risk categories via a regression discontinuity design, or RDD. These are quasi-experimental methods that are commonly used to estimate treatment effects within the econometrics and causal inference literature [5, 35, 79]. The key insight which enables this approach is that the discrete DEWS

refer to students who had a 0 in the "is White" column.

label, or risk category, is generated by thresholding a smoothly varying and continuous variable. This sharp discontinuity serves as a "natural experiment" amongst the subset of students whose relevant statistics lie close to the threshold.

In more detail, as part of its prediction pipeline, DEWS generates confidence intervals around the predicted probability of on-time graduation (i.e. the DEWS score). If the upper bound on this predicted probability is below a department-chosen threshold of $t^* = .785$, then students are assigned a high risk label.[5] On the other hand, if the lower bound is above t^*, then students are predicted to be at low risk. Lastly, if t^* is contained within the confidence interval, then students are assigned into the moderate risk category. The intuition behind the thresholding is that if the upper confidence bound on a student's probability of graduation is low, then students are unlikely to graduate on-time and should hence be prioritized into the high risk category. The opposite rationale is true for low risk predictions.

Regression discontinuity designs are motivated by the observation that risk categories (i.e., the treatment) for students whose predicted confidence bounds are close to the threshold are essentially random. By looking at students whose upper confidence bound is just around this threshold, we can infer the expected difference in the likelihood of on-time graduation that results from changing a student's DEWS label from moderate to high risk. Similarly, comparing students whose lower confidence interval is around t^* reveals the analogous impact of assigning students to the moderate versus low risk category.

To make this formal, we use notation from the potential outcomes framework. We let $Y(r) \in \{0, 1\}$ denote the indicator variable for on-time student graduation under assignment into predicted risk category r, where r can take values in the set {low, moderate, high}, and let Y denote the observed historical outcome. We use x to denote the vector of student features and $(\ell(x), u(x)) \in [0, 1]^2$ to denote the lower and, respectively, upper confidence bounds determined by the DEWS system for each student. We assume that the conditional expectations of outcomes are smooth, continuous functions of these confidence bounds:

Assumption 6.4.1. (smoothness of conditional expectations) For all values of r in the set {low, moderate, high} the functions $\mathbb{E}[Y(r) \mid \ell(x) = c]$ and $\mathbb{E}[Y(r) \mid u(x) = c]$ are twice continuously differentiable and smooth functions of c, for all values c in an open set containing t^*.

Under this mild technical condition, it is well-known that performing local linear regression around the cutoff value estimates the desired causal effects up to a small bias

[5] This particular $t^* = .785$ threshold was chosen by the state at the beginning of the program to approximately balance the number of students assigned into each DEWS category.

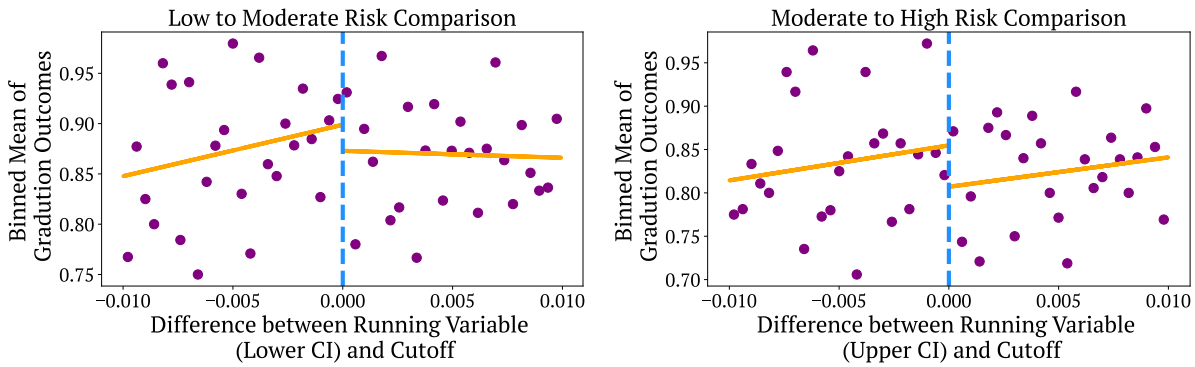


Figure 6.3: Visualization of RDD results. Purple dots correspond to the average graduation outcomes for students within each confidence interval bin. Yellow lines indicate fitted values from performing local linear regression. The gap between yellow lines at the point when they intersect the threshold (blue dotted line) is the point estimate for the treatment effect. **Left:** Treatment effect of increasing risk category from low to moderate. Students on the left side of the threshold are assigned to the moderate risk bucket. **Right:** Treatment effect of increasing risk category from moderate to high risk. Students on the left side of the threshold are assigned into the high risk bucket.

term. In particular, consider the treatment effect $\tau_{mod \to high}$ of increasing the DEWS label from moderate to high risk for students whose predicted upper confidence bound is equal to the threshold:

$$\tau_{mod \to high} := \mathbb{E}[Y(\text{high}) - Y(\text{moderate}) \mid u(x) = t^*].$$

This treatment effect is equal to the $\overline{\tau}_{mod \to high}$ that solves the ordinary least squares objective,

$$\mathbb{E}1\{|u(x) - t^*| \leq h\}(Y - \alpha - \overline{\tau}_{mod \to high} \cdot 1\{u(x) < t^*\} - \beta \cdot (u(x) - t^*) \cdot 1\{u(x) < t^*\} - \gamma \cdot (u(x) - t^*))^2,$$

up to bias that is order h^2 and statistical error that is order $n^{-1/2}$ [35]. Here, $h > 0$ is a bandwidth parameter which ensures that only points close to the cutoff enter the regression and n is the number of points within that bandwidth.

Likewise, if we let $\tau_{low \to mod}$ be the average difference in graduation rates resulting from increasing the predicted risk category from low to moderate,

$$\tau_{low \to mod} := \mathbb{E}[Y(\text{moderate}) - Y(\text{low}) \mid \ell(x) = t^*], \tag{6.1}$$

Causal Effect		Point Estimate	95% Confidence Interval	p-value	n
$\tau_{low \to mod}$:	increasing risk from low to moderate	0.026	(-0.024 0.076)	0.31	2653
$\tau_{mod \to high}$:	increasing risk from moderate to high	0.048	(-0.02, 0.116)	.17	1888

Table 6.1: Causal effect estimates from regression discontinuity design. The 95% confidence intervals are computed from using a Normal approximation. The n values correspond to the number of points that lie within the chosen bandwidth of the threshold. We cannot reject the null hypothesis that the predicted risk category has no effect on the average graduation outcome.

then this treatment effect is also $O(h^2 + n^{-1/2})$ close to the value $\overline{\tau}_{low \to mod}$ that solves:

$$\mathbb{E}1\{|\ell(x) - t^*| \leq h\}(Y - \alpha - \overline{\tau}_{low \to mod} \cdot 1\{\ell(x) < t^*\} - \beta \cdot (\ell(x) - t^*) \cdot 1\{\ell(x) < t^*\} - \gamma \cdot (\ell(x) - t^*))^2.$$

We present the results of this regression discontinuity analysis in Figure 6.3 and Table 6.1. Overall, we do not find strong enough evidence that assigning students into different risk categories changes their resulting probability of graduation. We estimate that increasing the DEWS category from moderate to high risk increases like likelihood of on-time graduation by less than 5%, while the analogous change from low to moderate risk leads to an even smaller increase of around 3%. In either case, we cannot reject the null hypothesis that the predicted DEWS category has no effect on graduation outcomes. The 95% confidence interval for the moderate to high risk treatment effect ranges from -2% to 11.6%. On the other hand, we can, with high probability, rule out that these changes in predicted risk have a very *large* impact on their eventual probability of on-time graduation. The upper confidence bounds for either regression indicate that it is unlikely that treatment effects are larger than 12% for the moderate to high risk comparison, or 8% for the low to moderate comparison.

Going back to our early warning systems as sorting algorithms analogy, the finding that DEWS predictions have no impact on graduation outcomes implies that it does not matter where students are placed in the sorted list. Their resulting probability of graduation is the same. While the list of students is accurately sorted according to dropout risk, the thresholding into discrete risk categories does not lead to effective interventions that lead to large changes in the likelihood of on-time graduation. This lack of impact on bottom-line outcomes means that it is quite unlikely that the differences in predictive performance amongst subgroups identified in Section 6.3 amplified, or in any way altered, existing high school graduation gaps between students in Wisconsin.

Robustness Checks. In both cases, we restrict our regression analysis to the population of students whose school districts log on at least once per year during the period for which we have usage data and choose the bandwidth parameter h to be .01. However, the high-level conclusions are robust to the choice of bandwidth parameter. Furthermore, similar conclusions are obtained if we restrict the analysis to just include the population of students of color, or students who qualify for free or reduced lunch, amongst others. Please see the supplementary material (Section 6.7) for a more comprehensive sensitivity and robustness analysis of the regression discontinuity design.

Further Analyses. The regression discontinuity design tries to infer whether the choice of risk category changes the likelihood of on-time graduation for a very particular subset students who lie just on the margin of being assigned either label. However, this is only one, of many possible ways in which DEWS predictions can impact outcomes. A necessary and sufficient condition for DEWS predictions to causally influence graduation outcomes is if there is any statistical dependence between these two variables, that is not explained in the features.

Drawing upon recent work by [56], in the supplementary material (Section 6.7), we present a different, "predicting from predictions" approach which directly tries to tackle this question. However, the conclusions from those experiments match those of the regression discontinuity design: there is no evidence that DEWS predictions have in any way influenced the likelihood of on-time graduation.

6.5 Why DEWS is Accurate, but Ineffective

As described in the introduction, the accuracy but non-impact of the DEWS system can be explained by a single, overarching empirical fact:

$$\textbf{Outcomes} \perp\!\!\!\perp \textbf{Individuals} \mid \textbf{Environment}$$

That is, observed graduation outcomes are (approximately) statistically independent of individual information, such as individual test scores or even demographic information, conditioned on students' academic and social environments.

We structure this section as follows. First, we develop the idea behind this conditional independence statement. Then, before presenting the empirical evidence that led us to this claim, we assume it holds, and outline how it explains the empirical findings on DEWS presented so far. Finally, at the end, we present a number of different experiments establishing how the major implications of the conditional independence statement are all empirically consistent with the historical data. This statistical law of public education

cannot be substantially refuted on the basis of a decade's worth of detailed statistics from the state's public schools.

We start by describing the social thesis in more detail. Recall that DEWS uses over 40 different features in order to predict a student's likelihood of on-time graduation. These range from basic demographic statistics like race and gender, to socioeconomic status (e.g., qualifying for free or reduced lunch), as well as a number of features that do not pertain to the individual student *per se*, but rather to their community (e.g., percent of their cohort that is non-white). Conceptually, we can take this large set of features and partition them into two main groups: environmental features and individual features.

We define a feature as individual if it measures information that directly corresponds to a specific student. Examples of these features are variables like race, gender, or the number of days a student attended school. On the other hand, we define environmental features to be those which capture information about a student's community. These are variables like the size of a student's cohort, the average math score in their school, or the median income in their district.

The thesis is a claim about observed historical patterns in the data. It states that conditioning on students belonging to a specific academic environment (i.e the full set of environmental features), their academic outcomes (the binary indicator of on-time graduation) is approximately statistically independent of the full vector of individual features. That is, if we already know the environmental features, additionally knowing the vector of individual features provides little additional information regarding the likelihood of on-time graduation.

Importantly, the thesis is *not* a causal statement. From the point of view of interventions, it does not rule out the fact that intervening on individuals can lead to drastic changes in their educational outcomes. It also does not imply that individual level interventions are any more, or less, effective than intervening on student's environments. It is only a claim about observed statistical patterns in the data. We return to this point in the discussion section.

Next, we pause to consider how it can explain the behavior we've observed so far within the DEWS system. As we presented in Section 6.3, DEWS predictions are highly accurate. However, if environmental features are strongly correlated with graduation outcomes, including these as part of the model should already lead to a highly accurate predictor. Recall that the DEWS model does indeed incorporate a large swath of features describing a students' communities (see the table in Section 6.7 for a full list), and these features vary significantly from district to district. Therefore, the independence thesis is consistent with the accuracy displayed by DEWS.

More importantly, the conditional independence thesis could explain why assigning students to different risk categories does not change their likelihood of on-time graduation

(recall the results from the RDD in Section 6.4). If outcomes are independent of individual features conditioned on the environment, *within each school*, DEWS predictions are essentially random and unrelated to the true likelihood of on-time graduation. Counselors and data teams examining the scores would notice that the DEWS predictions are not particularly accurate for their students. Assuming that counselors are confident in their ability to identify students at least as well as a random predictor, they would hence ignore the DEWS score. If a significant fraction of counselors log onto the DEWS system, but find that the system's predictions are largely uninformative and should be disregarded, it stands to reason that altering a student's risk score would not lead to a large change in their individual likelihood of graduation.

Note that this lack of within school signal does not contradict the fact that DEWS predictions are accurate when evaluated across the state as a whole. Because schools across the state have very different graduation rates, it can be simultaneously true that predictions within each school are uninformative, but predictions across the entire state (or for particular demographic groups) are very accurate.

Empirical Evaluations of the Independence Thesis

Given that the conditional independence of individuals and outcomes is a strong claim, in principle, it can also be easily refuted empirically. In this section, we examine several of its main implications and verify to what extent these are indeed consistent with the observed data. More specifically, we evaluate two main consequences of the independence thesis and find that they are by and large consistent with historical data.

First, the thesis implies that once we are aware of a student's environment, the likelihood of on-time graduation is uncorrelated with individual statistics like the number of days a student attended school, their disciplinary history, or even their race. From the point of view of prediction, including individual-level features in a predictive model that already uses environmental-level covariates does not lead to better accuracy.

Second, it suggests that within a particular school (environment), all students should have nearly identical probabilities of graduating from high-school on time. Any variation in individual student features (e.g personal test scores) *within* a particular district should not significantly change the probability that a student finishes school on time.

For these experiments, in addition to the features present within the DEWS system, we incorporate additional environmental features with the hopes of gaining a more complete understanding of exactly what predicts on-time graduation. In particular, we include community-level statistics regarding students' public school districts drawn from the US Census Bureau's American Community Survey and the National Center for Education Statistics. These span areas like school expenditures per student and racial demographics

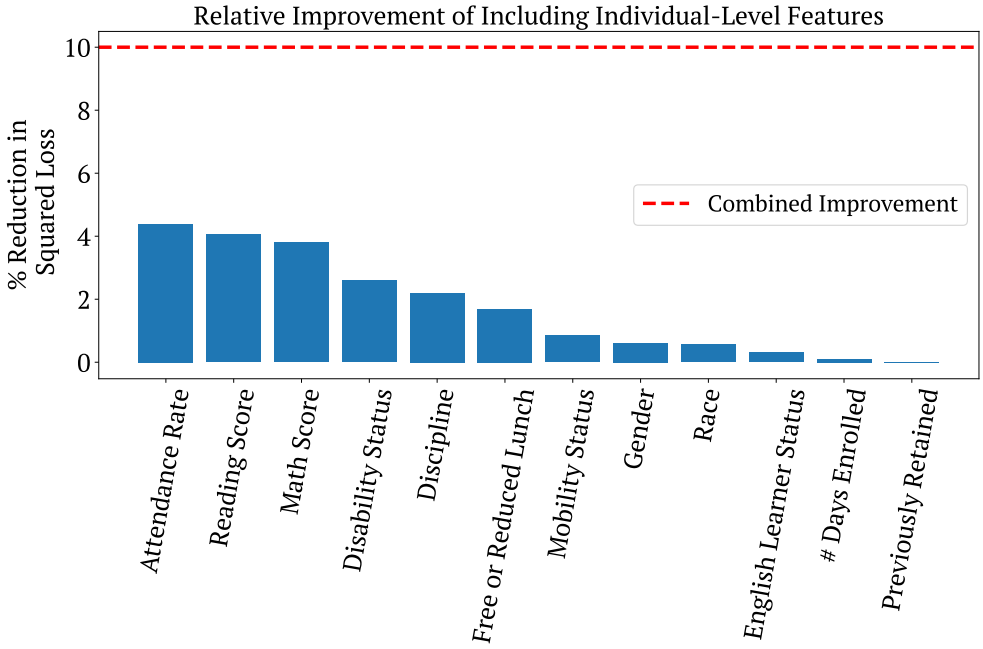


Figure 6.4: Relative improvement in predictive accuracy gained by incorporating individual-level features into a predictor that only uses environmental covariates. The red dotted line indicates the percent improvement in squared loss that is achieved by including all individual features. The blue bars denote the improvement achieved by adding a single category of individual features to the model.

in the district. Please see the supplementary material for a full list of features and their relevant categorizations.

Metric	Squared Loss	Log Loss	0-1 Loss	AUC
Absolute	.006	.03	.004	.09
Relative	10%	10%	6%	11%

Table 6.2: Relative improvement in performance achieved by including individual features on top on environmental features. For each metric, we report both the absolute improvement and the relative improvement, where relative is with respect to the performance of the environmental predictor.

Irrelevance of Individual Features for Prediction. We begin by testing the first claim that including individual-level features in a predictive model that already uses environmental-level covariates does not lead to better accuracy. Our statistical analysis closely relates to a work by [32] who motivate a similar distinction between individual features and background features in a prediction problem, and propose statistical methods to estimate the degree to which a predictor relies on background features.

For our experiment, we take the entire dataset of 8th graders for which graduation outcomes have been observed and assign 80% of them (≈160k) to a training set and 20% (≈40k) to a test set. On the training set, we fit two separate models: one model that forecasts graduation outcomes just using the environmental features and a different model that predicts on-time graduation using both individual features and environmental features. Both training procedures are identical learning algorithms (i.e gradient-boosted decision trees), use the same training data, and only differ in the subset of features that they have access to. We evaluate both of these models on the held-out test set in order to assess the marginal improvement achieved by including individual-level features.

In addition to analyzing the overall value of including the complete set of individual features, we also consider the partial benefit of adding including particular subsets of these features to an environmental-based predictor. In particular, we repeat the same training procedure as before and produce models that predict on-time graduation using all the environmental features plus a specific individual feature.

We present the results of these experiments in Figure 6.4 and Table 6.2. Adding individual features improves the predictability of on-time graduation by roughly 10% across a number of common metrics such as log loss, squared loss, or AUC. Amongst the individual features, the single most important variable is attendance rate which improves squared loss by 4%, followed by student's scores on standardized exams.[6]

A 10% reduction in squared loss is not zero, hence the independence statement does not hold exactly, but only approximately. Adding individual features leads to a marginally better prediction. To get a better perspective of exactly how much better, we can examine the calibration curves of both models (Figure 6.5).

As we can see, both the environmental predictor and the model that uses the entire set of features generate almost perfectly calibrated predictions. The main qualitative difference between these two is that the model that uses individual features does slightly better at identifying students with very low graduation rates. From looking at Figure 6.5, the predictor that includes individual features generates slightly better predictions near the bottom left of the calibration curve. However, this difference is relatively minor since

[6]These results are consistent with previous work on early warning systems that noted the predictive value of attendance rate on future dropout [4, 8].

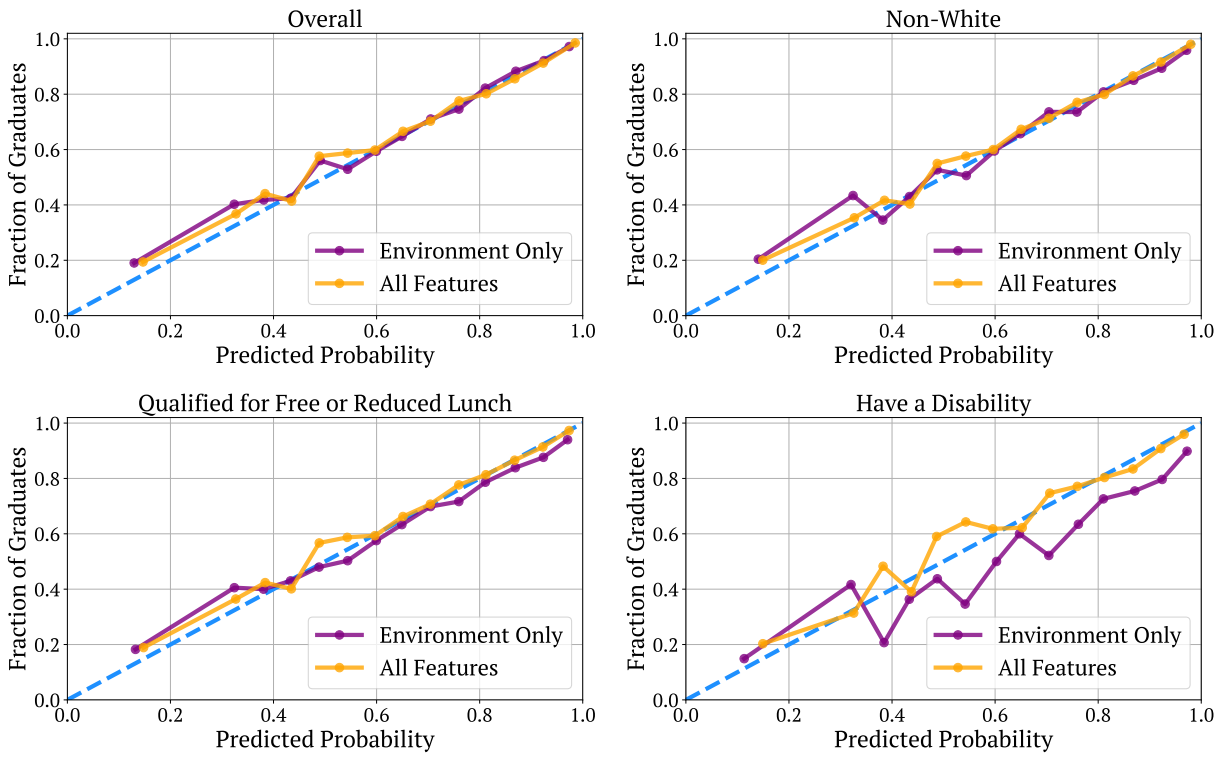


Figure 6.5: Comparing calibration curves for the environmental predictor versus the full-featured predictor, amongst different individually-defined student groups. **Top Left:** comparison between white versus students of color. **Top Right:** comparison between students who have been diagnosed with a disability versus those that have not. **Bottom Left:** comparison based on qualifying for free or reduced lunch. **Bottom Right:** calibration comparison between male and female students. Overall, student of color groups tend to have lower calibration error.

less than 2% of students have predicted probabilities of graduation less than 40%.[7]

Even more surprisingly, both models remain calibrated even when evaluated on important subgroups. In Figure 6.5, we see that the predictions for this model also lie on the $y = x$ diagonal when we restrict evaluation to students of color, or students that qualify for free or reduced lunch. This is quite impressive. Note that these subgroups are defined in terms of individual-level features. Therefore, the environmental predictor does *not* have access to this particular piece of important information when generating the prediction, yet its prediction is still well-calibrated for this group of students. Outcomes for these

[7]See Figure 6.11 in the supplementary material for the full histogram of predicted probabilities.

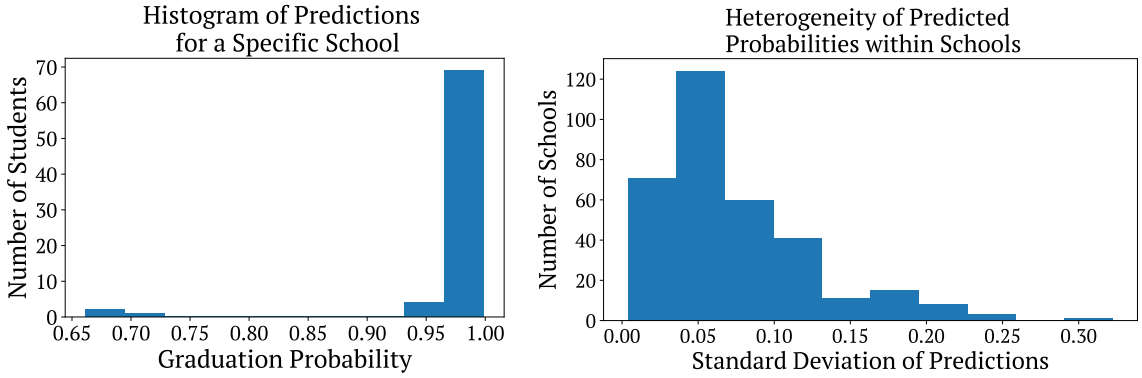


Figure 6.6: **Left:** Histogram of predicted probabilities generated by the model that uses both environmental and individual features for a students belonging specific school in the held-out test set. **Right:** Histogram of the standard deviation of predicted probabilities within each school across all schools in the test set with more than 30 students.

specific students are still statistically largely largely determined by their environments.

Same School $\implies$ Same Probability of Graduation. Naively, we might imagine that within each school, there are above average students who have very high probabilities of graduating on-time, a large number of people who graduate at about the school average, and a tail of struggling students that are likely to drop out. However, this phenomenon is ruled out under the independence thesis. The second implication we outlined previously is that all students in the same school have nearly identical probabilities of graduating, regardless of their individual features.

Empirically, this also turns out to be largely true. Schools are homogeneous populations composed of students that have nearly identical likelihoods of on-time graduation.

To see this latter point, on the left side of Figure 6.6, we plot the histogram of predictions made by the model which uses the entire set of available features (both individual and environmental features) for a specific school in the held-out test set. Even though the model has the ability to assign different predictions to different students, in practice, most students within this school have nearly identical predicted probabilities of graduation. Far from being an outlier, this particular histogram of within-schools predictions is representative of schools throughout the state. On the right side of Figure 6.6, we present the histogram of standard deviations of the within-school predictions made by this model

across all schools in the test set with at least 30 students.[8]

The main insight from this histogram is that the distribution of standard deviations is very concentrated on small values. For the vast majority of schools, most students within the school receive almost exactly the same predicted probability of graduation.

This lack of within-school variation is in line with the initial experiment discussed earlier demonstrating the relative similarities in predictive performance between environmental and individual predictors. If there were large amounts of variation in the true probabilities of graduation within each school, a model that uses informative individual features would significantly outperform the environmental predictor that outputs a constant prediction. [9]

This finding is also in line with the independence thesis which claims that within a particular environment, the likelihood of on-time graduation is essentially constant across students.

Same School, Same Individual Features. This last observation is not a direct consequence of the conditional independence statement. However, it provides further insight into the patterns observed within the DEWS system and the relative value of individual features.

We find that the very notion of an "individual feature" is somewhat of a red herring. Due to the high levels of socioeconomic and racial segregation between public school districts, students within a particular district tend to have very similar "individual features". If all students within the same school have the same individual features, then these features are largely meaningless for the sake of prediction. It becomes information-theoretically impossible to disambiguate different levels of dropout risk amongst students within the same school environment.

For example, one take away from the calibration plot in Figure 6.5 is that individual-level statistics such as race and socioeconomic status are so strongly correlated to the available environmental features, that it suffices to just know the environmental features in order to recover these individual features. Within the context of Wisconsin public schools, these two features are close to constant within a particular environment. This

[8]That is, we group students in the held out test set according to their school IDs and compute the standard deviation of the predicted probabilities for these students. The histogram displays the frequencies of these standard deviations across all schools.

[9]These conclusions are derived on the basis of the outputs on a predictor that was trained on finite amounts of data and used limited amounts of computation. Therefore, it is possible that the learning algorithm failed to pick up on existing variation in the true probabilities of graduation within specific schools. This is however, quite unlikely. Methods like gradient-boosted decision trees are widely believed to achieve Bayes' risk for similar social science datasets where the number of data points far exceeds the number of features.

explains why the model that uses only environmental features can generated calibrated predictions on subgroups defined in terms of individual-level features.

Further experiments confirm this view. On average 75% of the population in Wisconsin is non-White. However, a predictor that uses environmental features to predict whether a student is a person of color achieves a 0-1 error of .17, which is significantly better than random prediction (random guessing gets .25). Similarly, 37.5% of the students in our dataset qualify for free or reduced lunch. Yet, predicting free or reduced lunch status only using the environment achieves a 0-1 loss of .28. Random guessing would only achieve a misclassification error of .375 (lower numbers are better).

These experimental results are conducted in the same fashion as the previous ones. We fit a predictor on the training set that only uses environmental features. However, in this case the target variable is the indicator for being a student of color, or (in a separate experiment) qualifying for free or reduced lunch. The accuracy of these models is evaluated on the held out test set.

Statistically speaking, even things like test-scores are environmentally determined. Standardized exams are explicitly designed to generate a Gaussian-like distribution of scores that distinguishes between high performing students and low performing students. However, most of the variance in this distribution comes from the fact that students in different schools score very differently on the exam. Students within each school have very similar scores.

More formally, 80% of schools have a within-school variance of math test scores that is smaller than the state-wide variance in test scores. If schools were composed of identical sub-populations of students, the within-district variance would be smaller than the state average only 50% of the time. Furthermore, this is not true just for individual math scores, if we consider the full vector of individual-features, 78.5% of districts have lower variance than the state average.[10]

This relatively small variation of individual features within each school is in line with the independence thesis. It is quantitative evidence that students within each district are essentially identical to one other. There is lower than average diversity in terms of academic performance, race, or socioeconomic status within the typical public school.

[10]That is, if we denote the vector of individual features by x, the variance of the random vector $\mathbb{E}[\|x-\mathbb{E}x\|^2]$ is smaller when we take the expectation over x's to be over students in a specific school rather than the state as a whole.

6.6 Discussion: The Marginal Value of Prediction in Education

Summarizing the main empirical results of our work, we find that over the past decade the DEWS system has provided a highly accurate assessment of the risk that individual students in Wisconsin will leave high school without a degree. However, despite this high degree of accuracy, the availability of DEWS predictions has not translated into improved outcomes. Both of these findings are explained by the fact that individuals and graduation outcomes are largely independent statistically once we condition on students' school environments.

These findings have several direct implications regarding the use of machine learning in US public education and the design of early warning systems. Early warning systems are intended as tools to help school administrators efficiently allocate scarce resources amongst their students. In large districts, teachers and counselors cannot always spend significant amounts of time with each individual student. They are forced to make difficult choices regarding which students to help due to insufficient resources. Risk prediction tools like early warning systems are meant to help with this decision-making process by systematically ranking students according to their likelihood of leaving high school without a diploma.

Due to the influence of environmental factors and the degree of racial and socioeconomic segregation between school districts, the allocation suggested by an early warning system is not significantly better than random guessing. If we assume that teachers and counselors are already capable of allocating school resources at least as well as random guessing, the marginal value of creating an early warning system is near zero. Since students within each school have nearly identical probabilities of leaving high school without a diploma, there is not much to be gained from a fine-grained ranking of these students. Or, in more technical terms, there is no reason to optimize past the noise floor.

This does not mean that finding effective ways of allocating resources to schools is by any means hopeless. In fact, our analysis establishes that the following targeting strategy is near optimal: rank schools by graduation rate and intervene on the schools with the lowest rates until resources are expended. There are five high schools (out of a total of over 500) in Wisconsin whose graduation rates are close to 70%. This 70% number is about the same as the graduation rate amongst the students who were assigned a high-risk prediction by the DEWS system. Over two thirds of students in these schools are Black or Latino and over 90% qualify for free or reduced lunch. Taken together, these schools generate 10% of all students who leave high school without a diploma in the state. Raising graduation rates in these schools closer to the state average (90%) would significantly

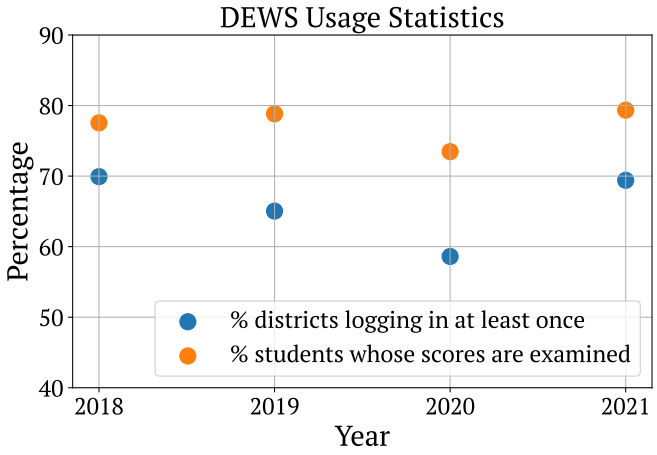


Figure 6.7: Visit statistics describing how often counselors from various public school districts in the state log onto the WISEDash platform to examine DEWS scores. In **blue**, we plot the fraction of school districts that log on at least once per year. In **orange**, we weight visits statistics by the fraction of total students in that district to approximately measure the fraction of students in the state whose DEWS scores are examined. See footnote for a formal description.

reduce the dropout rate.

Evidently, identifying a large subset of students that needs help is in many ways a trivial task. Not only can it be done without the use of individual student information, it can also be done without any machine learning at all. This school based strategy requires significantly less effort and technical sophistication than designing a well-functioning machine learning system. The relevant question here is not *who needs help?*, but rather *it's clear who needs help, what should we do about it?*.

6.7 Supplementary Material

Investigations into DEWS System Usage

Predictions that are never looked at can never influence outcomes. Hence, as part of our investigations into the impacts of the DEWS program, we first verified that the system had been actively used. As discussed in the overview, the DPI issues credentials that school administrators can use to log onto a department website and examine scores for students in their district. Each set of login credentials is associated with a unique district id. Using

this data, we were able to examine the number of times each district logs on to the DEWS platform each month for the period from 2018 to 2021. Due to software changes, data before 2018 is unfortunately not available.

In Figure 6.7, we plot in blue the fraction of districts that log on at least once per year during the time frame in consideration. Because districts can directly download all student scores after a single yearly visit, we opt to measure usage by whether districts log on at least once, rather than the total number of visits. Using this statistic, we find that around two thirds of districts regularly log to the WISEDash portal each year. Other than a drop during the start of the COVID-19 pandemic in 2020, there are no clear trends in usage across time. While DEWS may perhaps have seen lower usage in the period before 2018 as the system was less well-known, from this data we conclude that utilization has been relatively consistent across time.

To gain better intuition of where the system may have been most effective, we also evaluated whether there are any major differences between districts that regularly log on and those that do not. In the orange dots in Figure 6.7, we plot an estimate of the total fraction of students whose DEWS scores are examined. More specifically, we weight the indicator variable of a school district logging onto the DEWS platform by the fraction of the state's student population that is in that district.[11] Because we cannot disambiguate between users from the same district but different schools, this statistic can overstate the true fraction of students whose DEWS scores are acted upon. This bias is relatively minor, however, since most districts have just one public school in them. With this caveat in mind, our analysis shows that the larger districts use DEWS more often. This observation is consistent with the fact that most of Wisconsin consists of small rural districts where teachers get to know students personally. Schools in these areas have less use for a system like DEWS.

Furthermore, using data from the American Community Survey, in Table 6.3 we look

[11] If we let $d_{i,j}$ denote the indicator variable for district i logging on during year j and n_j denote the the number of students in district j (which is roughly constant across years), in orange we plot $(\sum_i d_{i,j} n_i)/\sum_i n_i$ across all years j.

	Median Household Income	% Black	Total Population	Graduation Rate
% Years Visited	-.01	.08	.20	-.16

Table 6.3: Correlation between the fraction of years a school district visited the WISEDash portal between 2018 to 2021 and district-level socioeconomic statistics drawn from the ACS 5 year community survey.

at the correlation between the fraction of years for which a district logs on at least once and other socioeconomic information from that district. From this, we see that districts with lower graduation rates and higher poverty indices tend to log on most often. Summarizing, the DEWS system has been actively examined by the majority of public schools in Wisconsin. Furthermore, districts with the highest need also have the highest usage.

Description of System Features

In this subsection, we describe the entire set of features we use in our experiments and analysis of the DEWS system. We divide these features according to the source they are drawn from: DEWS, the 5-year American Community Survey, and the National Center for Educational Statistics.

DEWS Features

Below, we list all features included in the DEWS system. In addition to providing a short description, we indicate the relevant partition it belongs to for the comparisons between different feature sets discussed in Section 6.5.

Feature	Partition	Description
Gender	Individual, Non-Malleable	Student's gender, can be male or female
Race	Individual, Non-Malleable	Race is coded as a combination of 7 mutually exclusive indicator variables for whether student is White, Black, Hispanic, Native American, Pacific Islander, Asian, or belongs to two or more races
Disability Status	Individual, Non-Malleable	Indicator for whether student has been identified as having a disability. See DPI website for background on disability classifications. Some classifications such as the Emotional Behavioral Disability, are strongly correlated with race and socioeconomic status
English Learner Status	Individual, Malleable	Indicator of whether English is the student's native language. If the student is a non-native speaker, there are additional features describing whether English skills are low, moderate, or high
Free or Reduced Lunch Status	Individual, Non-Malleable	Indicator variable for whether the student's household income is above or below a certain threshold based off of the federal poverty line. This threshold is

Feature	Partition	Description
Enrolled Days	Individual, Malleable	Number of days a student has been enrolled in school over the last year
Reading Score	Individual, Malleable	Reading score on state-wide standardized exam
Math Score	Individual, Malleable	Math score on state-wide standardized exam
Full Academic Year - District	Individual, Malleable	Indicator variable for whether student has been in the current school district for the entire academic year
Full Academic Year - School	Individual, Malleable	Indicator of whether student has been in the current school for the entire academic year
Disciplinary Incidents Count	Individual, Malleable	Number of disciplinary incidents over the previous year
Days Removed	Individual, Malleable	Number of days suspended from school
Removal Type	Individual, Malleable	Indicator of whether student was expelled or suspended from school
Disciplinary Descriptors	Individual, Malleable	Separate indicator for whether the disciplinary incident was assault, drug related, involving a weapon, or other
School Count	Individual, Malleable	Number of unique schools enrolled in during the last year
District Count	Individual, Malleable	Number of unique school districts enrolled in during the last year
Enrollment Count	Individual, Malleable	Number of enrollment spells during the last year

Feature	Partition	Description
Cohort Reading Scores	Environmental, Non-Malleable	Mean and standard deviation of student's school cohort reading scores on state exams
Cohort Math Scores	Environmental, Non-Malleable	Mean and standard deviation of student's school cohort math scores on state exams
Cohort Size	Environmental, Non-Malleable	Number of students in cohort
Cohort Suspended	Environmental, Non-Malleable	Number of peers in cohort who have at least one suspension
% of Cohort with a Disability	Environmental, Non-Malleable	Percentage of peers in student's cohort that have a disability
% of Cohort FRL.	Environmental, Non-Malleable	Fraction of cohort qualified for free or reduced lunch
% of Cohort Non-White	Environmental, Non-Malleable	Fraction of student's cohort that is non-White
Cohort Attendance	Environmental, Non-Malleable	Mean and standard deviation of attendance rate for cohort

Features from the American Community Survey

We downloaded this data from the 2015 5-year American Community Survey using the censusdata Python package. Data is aggregated at the public school district level and matched to students via district IDs provided by the department. While student features

are associated with a specific school year, we perform a many to one mapping and assign all students from the same district (and across all available school years starting in 2013-14) to the same ACS data. These community level statistics are, however, stable across time hence the mismatch in years is relatively minor.

We list the ACS variable codes and names below. The interested reader can visit the ACS website for a more comprehensive description regarding how these variables are defined and measured.

Variable Code	Name
DP02_0006P	Percent of families with male householder, no wife present
DP02_0008PE	Percent of families with female householder, no husband present
DP02_0151PE	Percent of total households with a computer
DP05_0001E	Estimate of total population
DP02_0152P	Percent of total households with a broadband Internet subscription
DP02_0059PE	Percent of population 25 years or older who did not complete 9th grade
DP02_0060PE	Percent of population 25 years or older who completed 9th grade
DP02_0061PE	Percent of population 25 years or older with a high school degree
DP02_0062PE	Percent of population 25 years or older who attended some college, but did not graduate
DP02_0063PE	Percent of population 25 years or older with an associate's degree
DP02_0064PE	Percent of population 25 years or older with a bachelor's degree
DP02_0065PE	Percent of population 25 years or older with a postgraduate degree
DP02_0066PE	Percent of population who completed high school or higher degree
DP02_0067PE	Percent of population who has a bachelor's degree or higher degree
DP02_0079PE	Percent of population living in the same house as a year before
DP02_0087PE	Percent of total population that is native born
DP02_0092PE	Percent of total population that is foreign born
DP03_0005PE	Unemployment rate for those 16 years of age or older
DP03_0062E	Median household income (2015 inflation-adjusted dollars)
DP03_0096PE	Percent with health insurance coverage
DP03_0119PE	Percentage of families whose income is below the poverty level
DP05_0017E	Median age
DP05_0032PE	Percent white
DP05_0033PE	Percent Black or African American
DP05_0039PE	Percent Asian
DP05_0066PE	Percent Hispanic or Latino (of any race)

Features from the National Center for Education Statistcs

The National Center for Education Statistics maintains an online portal called the Elementary / Secondary Information System (ELSI) whereby one can access financial data for public school districts in the US. Using their website, we pull the following set of features, each of which is associated with particular district ID and school year. This allows us to match the statistics with the dataset of individual student records. Please see the ELSI webiste for a full glossary of terms. All of these features fall into the environmental partition.

Variable Name
Total Current Expenditures: Other El-Sec Programs per Pupil[12]
Total Current Expenditures: Salary per Pupil
Total Revenue - Federal Sources per Pupil
Total Revenue per Pupil
Total Expenditures - Capital Outlay per Pupil
Total Revenue - Local Sources per Pupil
Total Expenditures per Pupil
Total Current Expenditures - Non El-Sec Programs per Pupil
Total Current Expenditures - Instruction per Pupil
Total Current Expenditures per Pupil
Total Current Expenditures - Support Services per Pupil
Total Revenue - State Sources per Pupil
Instructional Expenditures per Pupil
Total Current Expenditures - Benefits per Pupil

[12] El-sec expenditures refers to total current expenditures for public elementary and secondary education that are associated with the day-to-day operations of the school district.

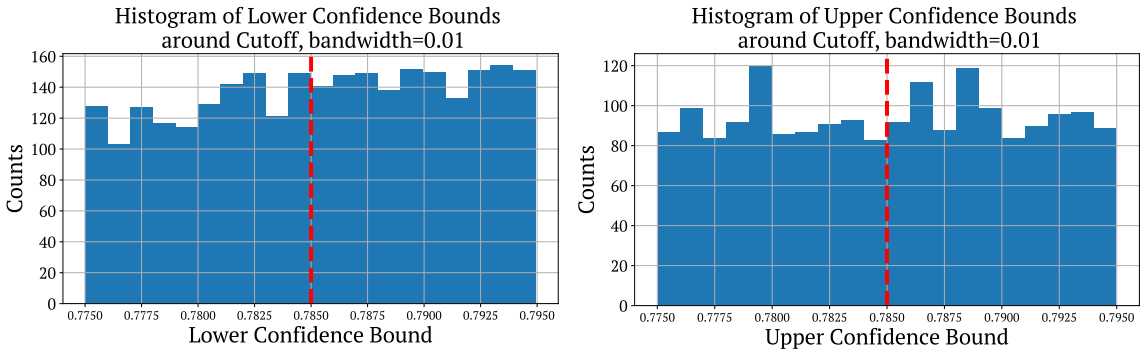


Figure 6.8: Histograms of the predicted confidence intervals (running variables) around the threshold for both RDD analyses. **Left:** lower confidence bound histogram for low versus moderate comparison. **Right:** upper confidence bound histogram for moderate versus high risk comparison.

Supporting Analysis: Do DEWS Predictions Lead to Better Graduation Outcomes?

In this section of the supplementary material, we present additional analyses and robustness checks supporting the claims presented in Section 6.4.

Regression Discontinuity Design: Robustness Checks

Inspecting Manipulation near Threshold A core assumption enabling regression discontinuity designs is that treatment assignment (in our case, the predicted risk category) is essentially random near the threshold. For a small choice of bandwidth, we should therefore observe that the histogram of the running variable (the upper or lower confidence bounds) on which we regress outcomes, is approximately uniform around the cutoff point.

As illustrated in Figure 6.8, we find that this is indeed the case. There is no evidence of "bunching " or manipulation for the population of students whose data is included in the regression analysis. This uniformity of scores is consistent with the idea that treatment (predicted DEWS label) is essentially determined by random assignment near the threshold.

Evaluating Covariate Balance In addition to checking for bunching, a different way of assessing whether there is indeed natural variation around the threshold is to compare

Causal Effect	Bandwidth h	Point Estimate	95% Confidence Interval	p-value	n
increasing risk from low to moderate	.005	0.0302	(-0.04 0.01)	0.39	1390
	.02	.0185	(-.017, .055)	.313	5264
increasing risk from moderate to high	.005	0.0574	(-0.04,.155)	.25	952
	.02	.0239	(-.025, .073)	.337	3461

Table 6.4: Sensitivity of regression discontinuity design to choice of bandwidth. We find that the conclusions derived from the regression analysis are largely insensitive to the choice of bandwidth h. Doubling or halving the bandwidth does not significantly the results presented in Table 6.1.

the characteristics of students that fall within the bandwidth.

More specifically, we can verify that important covariates are balanced around the threshold. In Figure 6.10, we plot the results of this comparison. We find that students falling above or below the t^* threshold have very similar characteristics: they have similar attendance rates, demographic characteristics, and environmental features, amongst others. These small differences in covariates are again consistent with the idea that there is natural variation around the treatment threshold.

Sensitivity to Bandwidth Choice In the regression analysis presented in the main body of the paper, we chose the bandwidth parameter h to be .01. That is, we only included students if their upper (or respectively, lower) confidence bounds where within .01 of the $t^* = .785$ threshold. As discussed previously, the bandwidth parameter must be small to ensure a consistent estimate of the the treatment effect. Yet, the exact choice "how small" is somewhat arbitrary.

In Table 6.4, we present the results of performing the same regression analyses, but with different bandwidth parameters. We find that halving ($h = .005$), or doubling ($h = .02$), the bandwidth parameter, leads to qualitatively similar conclusions. There is no evidence that the predicted DEWS category causally influences graduation outcomes. In short, the high-level conclusions of the regression are stable under reasonable choices of bandwidth parameters.

Understanding Effects on Subgroups So far, we have studied what is the treatment effect of assigning students into different risk categories *on average* over the entire population of students. For the sake of completeness, we also assess whether the treatment effect of assigning students into different risk buckets is nonzero if instead of averaging over

Causal Effect	Subgroup	Point Estimate	95% Confidence Interval	p-value	n
increasing risk from low to moderate	Female	0.001	(-0.075, 0.078)	0.97	1133
	Non White	.0124	(-.058, .082)	.73	1298
	Free or Reduced Lunch	.0444	(-.013,.102)	.131	2078
increasing risk from moderate to high	Female	-.0007	(-0.1,.1)	.989	782
	Non White	.0184	(-.07, .107)	.683	1086
	Free or Reduced Lunch	.0281	(-.05, .106)	.480	1476

Table 6.5: Evaluating Causal Effects of Prediction on Subgroups. The table contains the results of rerunning the regression discontinuity design analysis presented in Section 6.4, but only considering students with particular features. The number of data points included in the regression is therefore smaller than that presented in Table 6.1. The main conclusions are, however, identical. We find no evidence that increasing the level of predict risk leads to a higher likelihood of graduation.

the entire population of students we restrict ourselves to looking at specific demographic groups. If treatment effects are heterogeneous, it is in principle possible for the effects to be zero on average over the entire population, but large over particular subgroups.

In Table 6.5, we present the results of running the regression discontinuity analysis where the population of students is further restricted to specific demographic groups: women, students of color (non-White), or students who qualify for free or reduced lunch. We find no differences in the influence of prediction on outcomes over these specific groups relative to the population as a whole. There is no evidence that the assigned risk category in any way influences the likelihood of on-time graduation.

Assessing Statistical Independence between Predictions and Outcomes

Our regression discontinuity design analysis centers on the impact of the predicted DEWS label on the likelihood of graduation for a particular subset of students whose confidence intervals lie at a specific threshold. While this effect is close to zero, one might still wonder whether DEWS predictions are influencing graduation outcomes ways that may not be captured by the RDD.

If we denote the vector of student features by X, the DEWS outputs (both the label and score) by $\hat{\mathcal{Y}}$, and the indicator variable of on-time graduation by Y, a necessary and sufficient condition for predictions to impact outcomes is that predictions and outcomes are not statistically independent given the features: $\hat{\mathcal{Y}} \not\perp\!\!\!\perp Y \mid X$. In other words, there is a specific subset of students with features X such that the choice of prediction $\hat{\mathcal{Y}}$ changes the dis-

tribution over outcomes Y. We can represent these independence statements graphically as seen in Figure 6.9.

Note that if the treatment effects estimated via the RDD are nonzero, then this disproves the conditional independence statement: predictions do change outcomes. However, the converse is not true. Predictions and outcomes could be statistically dependent even if the specific treatment effect estimated by the RDD is zero.

In this subsection, we directly tackle this question of detecting whether predictions and outcomes are statistically dependent. To disprove a conditional independence statement, it suffices to show that a predictor which predicts Y given X and $\hat{\mathcal{Y}}$ achieves higher predictive performance than one that just uses the vector X. Intuitively, higher predictive performance indicates that there is information about Y in $\hat{\mathcal{Y}}$ that is not fully contained in the features X. In order to identify dependence from this "predicting from predictions" approach it is necessary for the predictions $\hat{\mathcal{Y}}$ to be randomized functions of X.[13]

While the specific DEWS model for each year generates predictions $\hat{\mathcal{Y}}$ as deterministic functions of X, there is slight variation in models across various years. Recall that models are trained on a sliding window of the 5 most recent cohorts of students. Because training sets differ year over year, the resulting models are also different. We treat this variation between models as a natural source of randomness that enables this statistical independence test.

Using the full dataset for 8th grade predictions described at the beginning of Section 6.3, we perform an 80/20 train-test split as in the environmental vs. individual comparison from Section 6.5. On the training set, we generate two binary prediction models, one that predicts on-time graduation variable Y using the DEWS features X and the predictions

[13]See [56] for a more complete derivation of this approach and discussion of the relevant identifiability conditions.

Model	Squared Loss	Log Loss	0-1 Loss	AUC	r^2
Performative	0.052, (0.050, 0.054)	.185, (.180, .190)	0.065, (.063, .068)	.863	.184
Non-Performative	0.052, (0.050, 0.054)	.185, (.180, .191)	0.066, (.063, .068)	.862	.183

Table 6.6: Results for the statistical independence test comparing the predictive performance of performative and non-performative models. Entries represent point estimates and 95% confidence intervals derived from evaluating predictors on the test set. Confidence intervals are computed by bootstrapping. The identical performance of both models further illustrates how there is no evidence of predictions influencing outcomes.

$\hat{\mathcal{Y}}$ (DEWS score and label), versus one that just uses the features X. Following the nomenclature from [65], we refer to the model that includes DEWS outputs as additional covariates as the "performative" model. The predictions we train are again ensembles of state-of-the-art supervised learning methods for tabular data such as gradient boosted decision trees.

If there is no relationship between outcomes and predictions beyond that which is captured by the features, then the predictive performance of these performative and non-performative predictors should be identical on the held-out test set. This is exactly the results we observe in the experiment. From Table 6.6, we see that the "performative" model that includes DEWS predictions as features, has a statistically identical performance to a model that just uses the covariates X across a wide range of prediction metrics. These findings match those of the regression discontinuity design analysis: there is no evidence that DEWS predictions have in any way impacted on-time graduation. There is no stastistical evidence that DEWS data follows the causal model appearing on the right in Figure 6.9.

Figure 6.9: Causal diagrams illustrating possible relationships between features X, predictions $\hat{\mathcal{Y}}$, and outcomes Y in the DEWS system. **Left:** Predictions do not influence outcomes: $\hat{\mathcal{Y}} \perp\!\!\!\perp Y \mid X$. **Right:** Predictions change outcomes: $\hat{\mathcal{Y}} \not\perp\!\!\!\perp Y \mid X$. Because features are collected before outcomes Y are observed, and since predictions are generated on the basis of features, there are causal arrows $X \to \hat{\mathcal{Y}}$ and $X \to \hat{\mathcal{Y}}$. An effective early warning system should have data which follows the causal model on the right.

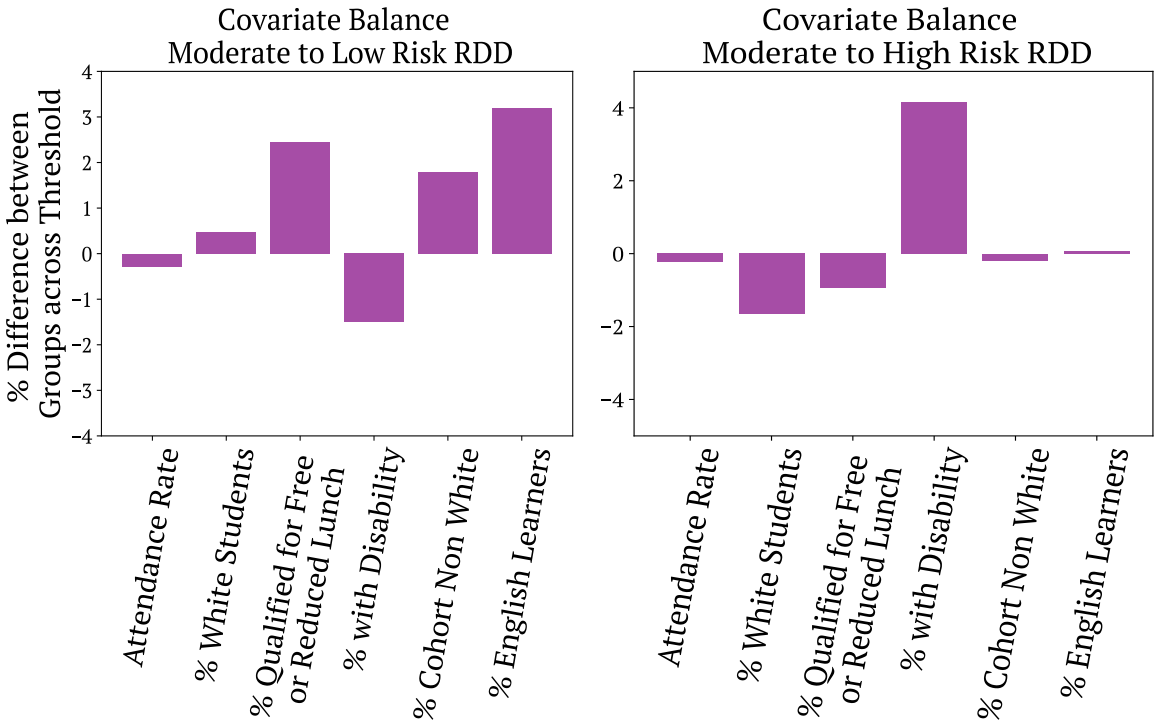


Figure 6.10: Comparing the features of students on either side of the RDD threshold. **Left:** Absolute difference between the average features for students on the left and right of the threshold for the moderate to low risk treatment effect RDD analysis. **Right:** Analogous comparison for the moderate to high risk RDD.Overall, we find that students on either side of the threshold have similar features.

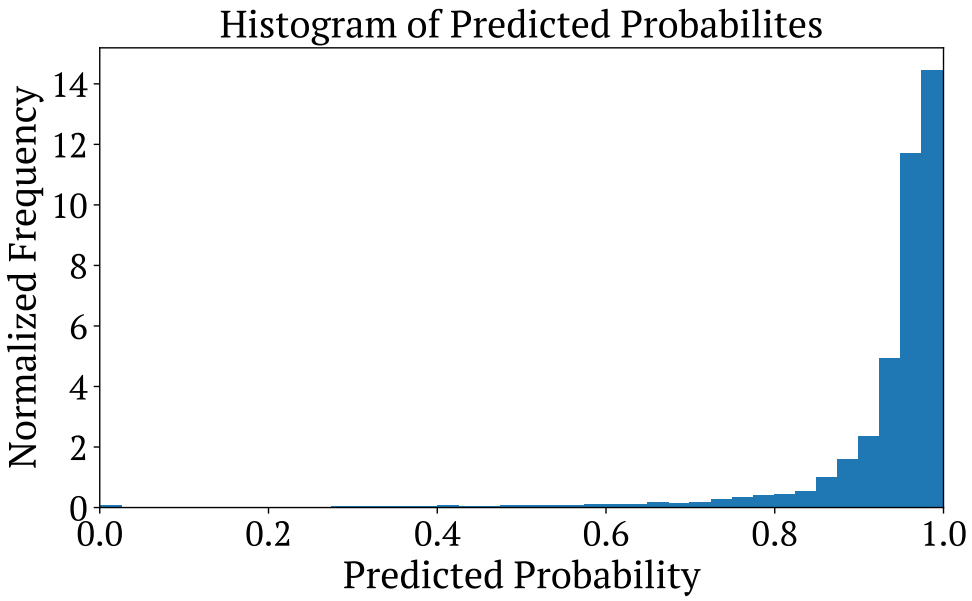


Figure 6.11: Histogram of predicted probabilities on the test set generated by the model from the experiments presented in Section 6.5 that uses the complete set of available features. The predictions from this model are highly calibrated as seen in Figure 6.5. 10% of students have predicted probabilities higher than .987 (that is, .987 is the 90% quantile). 10% of students have predicted probabilities *lower* than .85 (i.e. .85 is the 10% quantile). The vast majority of students have predicted probabilities of graduation between .9 and 1.

Bibliography

[1] Naoki Abe, Bianca Zadrozny, and John Langford. "An iterative method for multi-class cost-sensitive learning". In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004, pp. 3–11.

[2] Alekh Agarwal and Ofer Dekel. "Optimal algorithms for online Convex Optimization with multi-point bandit feedback". In: *Conference on Learning Theory*. 2010, pp. 28–40.

[3] Alekh Agarwal et al. "Taming the monster: A fast and simple algorithm for contextual bandits". In: *International Conference on Machine Learning*. PMLR. 2014, pp. 1638–1646.

[4] Elaine M Allensworth and John Q Easton. "What matters for staying on-track and graduating in chicago public high schools: A close look at course grades, failures, and attendance in the freshman year". In: *Consortium on Chicago School Research* (2007).

[5] Joshua D Angrist and Victor Lavy. "Using Maimonides' rule to estimate the effect of class size on scholastic achievement". In: *The Quarterly journal of economics* 114.2 (1999), pp. 533–575.

[6] John Langshaw Austin. *How to do things with words*. Oxford University Press, 1975.

[7] Robert Balfanz and Vaughan Byrnes. "Early warning indicators and intervention systems: State of the field". In: *Handbook of student engagement interventions* (2019), pp. 45–55.

[8] Robert Balfanz, Liza Herzog, and Douglas J Mac Iver. "Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions". In: *Educational Psychologist* 42.4 (2007), pp. 223–235.

[9] Alina Beygelzimer, John Langford, and Pradeep Ravikumar. "Error-correcting tournaments". In: *International Conference on Algorithmic Learning Theory*. Springer. 2009, pp. 247–262.

[10] Léon Bottou, Frank E Curtis, and Jorge Nocedal. "Optimization methods for large-scale machine learning". In: *SIAM Review* 60.2 (2018), pp. 223–311.

[11] Gavin Brown, Shlomi Hod, and Iden Kalemaj. "Performative prediction in a stateful world". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 6045–6061.

[12] Mary Bruce et al. "On track for success: The use of early warning indicator and intervention systems to build a grad nation." In: *Civic Enterprises* (2011).

[13] Michael Brückner, Christian Kanzow, and Tobias Scheffer. "Static prediction games for adversarial learning problems". In: *Journal of Machine Learning Research* 13.Sep (2012), pp. 2617–2654.

[14] Austin Derrow-Pinion et al. "Eta prediction with graph neural networks in google maps". In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2021, pp. 3767–3776.

[15] Miroslav Dudik et al. "Efficient optimal learning for contextual bandits". In: *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*. UAI'11. Barcelona, Spain: AUAI Press, 2011, pp. 169–178. ISBN: 9780974903972.

[16] Cynthia Dwork et al. "Calibrating noise to sensitivity in private data analysis". In: *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer. 2006, pp. 265–284.

[17] Cynthia Dwork et al. "Fairness through awareness". In: *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012, pp. 214–226.

[18] Cynthia Dwork et al. "Outcome indistinguishability". In: *ACM Symposium on Theory of Computing*. 2021. URL: https://arxiv.org/abs/2011.13426.

[19] Itay Eilat and Nir Rosenfeld. "Performative Recommendation: Diversifying Content via Strategic Incentives". In: *arXiv preprint arXiv:2302.04336* (2023).

[20] Charles Elkan. "The foundations of cost-sensitive learning". In: *International joint conference on artificial intelligence*. Vol. 17. Lawrence Erlbaum Associates Ltd. 2001, pp. 973–978.

[21] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[22] Clara Fannjiang and Jennifer Listgarten. "Autofocused oracles for model-based design". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 12945–12956.

[23] Clara Fannjiang et al. "Conformal prediction under feedback covariate shift for biomolecular design". In: *Proceedings of the National Academy of Sciences* 119.43 (2022), e2204569119.

[24] Ann-Marie Faria et al. "Getting students on track for graduation: Impacts of the early warning intervention and monitoring system after one year." In: *Regional Educational Laboratory Midwest* (2017).

[25] Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. "Online convex optimization in the bandit setting: gradient descent without a gradient". In: *Symposium on Discrete Algorithms*. 2005, pp. 385–394.

[26] Alex Frankel and Navin Kartik. "Improving information from manipulable data". In: *Journal of the European Economic Association* (2021).

[27] Oscar H Gandy. "Engaging rational discrimination: exploring reasons for placing regulatory constraints on decision support systems". In: *Ethics and Information Technology* 12 (2010), pp. 29–42.

[28] Parikshit Gopalan et al. "Loss minimization through the lens of outcome indistinguishability". In: *ITCS*. 2023.

[29] Parikshit Gopalan et al. "Omnipredictors". In: *ITCS*. 2022.

[30] Nika Haghtalab et al. "Maximizing welfare with incentive-aware evaluation mechanisms". In: *International Joint Conference on Artificial Intelligence*. 2020.

[31] Bernard E Harcourt. *Against prediction: Profiling, policing, and punishing in an actuarial age*. University of Chicago Press, 2006.

[32] Moritz Hardt and Michael P Kim. "Backward baselines: Is your model predicting the past?" In: *arXiv preprint arXiv:2206.11673* (2022).

[33] Moritz Hardt et al. "Strategic classification". In: *Proceedings of the ACM Conference on Innovations in Theoretical Computer Science*. 2016, pp. 111–122.

[34] Ursula Hébert-Johnson et al. "Multicalibration: Calibration for the (computationally-identifiable) masses". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 1939–1948.

[35] Guido W Imbens and Thomas Lemieux. "Regression discontinuity designs: A guide to practice". In: *Journal of econometrics* 142.2 (2008), pp. 615–635.

[36] Zachary Izzo, Lexing Ying, and James Zou. "How to learn when data reacts to your model: performative gradient descent". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 4641–4650.

[37] Zachary Izzo, James Zou, and Lexing Ying. "How to learn when data gradually reacts to your model". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 3998–4035.

[38] Zachary Izzo, James Zou, and Lexing Ying. "How to learn when data gradually reacts to your model". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 3998–4035.

[39] Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. "Regret minimization with performative feedback". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 9760–9785.

[40] Kaggle. *Give Me Some Credit*. `https://www.kaggle.com/c/GiveMeSomeCredit/data`. 2012.

[41] Michael P. Kim, Amirata Ghorbani, and James Zou. "Multiaccuracy: Black-box post-processing for fairness in classification". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 247–254.

[42] Michael P. Kim and Juan C. Perdomo. "Making decisions Under outcome performativity". In: *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*. Ed. by Yael Tauman Kalai. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023.

[43] Michael P. Kim et al. "Universal adaptability: Target-independent inference that competes with propensity scoring". In: *Proceedings of the National Academy of Sciences* 119.4 (2022), e2108097119.

[44] Jared E Knowles. "Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin." In: *Journal of Educational Data Mining* 7.3 (2015), pp. 18–67.

[45] Akshay Krishnamurthy et al. "Active learning for cost-sensitive classification". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1915–1924.

[46] John Langford and Alina Beygelzimer. "Sensitive error correcting output codes". In: *International Conference on Computational Learning Theory*. Springer. 2005, pp. 158–172.

[47] Johann Lau. *Google maps 101: How AI helps predict traffic and determine routes*. `https://blog.google/products/maps/google-maps-101-how-ai-helps-predict-traffic-and-determine-routes/`.

[48] Qiang Li, Chung-Yiu Yau, and Hoi-To Wai. "Multi-agent performative prediction with greedy deployment and consensus seeking agents". In: *arXiv:2209.03811* (2022).

[49] Martha Abele Mac Iver et al. "An efficacy study of a ninth-grade early warning indicator intervention". In: *Journal of Research on Educational Effectiveness* 12.3 (2019), pp. 363–390.

[50] Jane Macfarlane. "Your navigation app is making traffic unmanageable". In: *IEEE Spectrum: Technology, Engineering, and Science News* (2019).

[51] Donald A MacKenzie, Fabian Muniesa, Lucia Siu, et al. *Do economists make markets?: on the performativity of economics*. Princeton University Press, 2007.

[52] Nikhil Malik. "Does machine learning amplify pricing errors in housing market?: Economics of ml feedback loops". In: *Economics of ML Feedback Loops (September 18, 2020)* (2020).

[53] Debmalya Mandal, Stelios Triantafyllou, and Goran Radanovic. "Performative reinforcement learning". In: *arXiv preprint arXiv:2207.00046* (2022).

[54] Llew Mason et al. "Boosting algorithms as gradient descent". In: *Advances in Neural Information Processing Systems* 12 (1999).

[55] Nikolai Matni and Stephen Tu. "A tutorial on concentration bounds for system identification". In: *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE. 2019, pp. 3741–3749.

[56] Celestine Mendler-Dünner, Frances Ding, and Yixin Wang. "Anticipating performativity by predicting from predictions". In: *Advances in Neural Information Processing Systems*. 2022.

[57] Celestine Mendler-Dünner et al. "Stochastic optimization for performative prediction". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 4929–4939.

[58] John Miller et al. *WhyNot*. 2020. DOI: 10.5281/zenodo.3875775. URL: https://doi.org/10.5281/zenodo.3875775.

[59] John P Miller, Juan C Perdomo, and Tijana Zrnic. "Outside the echo chamber: Optimizing the performative risk". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 7710–7720.

[60] Smitha Milli et al. "The social cost of strategic classification". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*. Association for Computing Machinery, 2019, pp. 230–239. ISBN: 9781450361255.

[61] Mehrnaz Mofakhami, Ioannis Mitliagkas, and Gauthier Gidel. "Performative prediction with neural networks". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 11079–11093.

[62] Alfred Müller and Ludger Rüschendorf. "On the optimal stopping values induced by general dependence structures". In: *Journal of applied probability* (2001), pp. 672–684.

[63] Alfred Müller and Dietrich Stoyan. *Comparison methods for stochastic models and risks*. Vol. 389. Wiley, 2002.

[64] Adhyyan Narang et al. "Multiplayer performative prediction: Learning in decision-dependent games". In: (2022). URL: https://arxiv.org/abs/2201.03398.

[65] Juan Perdomo et al. "Performative prediction". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 7599–7609.

[66] Florian Pfisterer et al. "mcboost: Multi-calibration boosting for R". In: *Journal of Open Source Software* 6.64 (2021), p. 3453.

[67] Georgios Piliouras and Fang-Yi Yu. "Multi-agent performative prediction: From global stability and optimality to chaos". In: *arXiv preprint arXiv:2201.10483* (2022).

[68] Liudmila Prokhorenkova et al. "CatBoost: unbiased boosting with categorical features". In: *Advances in neural information processing systems* 31 (2018).

[69] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. "Making gradient descent optimal for strongly convex stochastic optimization". In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2012, pp. 1571–1578.

[70] Mitas Ray et al. "Decision-dependent risk minimization in geometrically decaying dynamic environments". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 7. 2022, pp. 8081–8088.

[71] Horn Roger and R Johnson Charles. *Topics in matrix analysis*. 1994.

[72] Sheldon M Ross et al. *Stochastic processes*. Vol. 2. Wiley New York, 1996.

[73] Matthew J Salganik et al. "Measuring the predictability of life outcomes with a scientific mass collaboration". In: *Proceedings of the National Academy of Sciences* 117.15 (2020), pp. 8398–8403.

[74] Ernst Friedrich Schumacher. *Small is beautiful: A study of economics as if people mattered*. Random House, 2011.

[75] Moshe Shaked and J George Shanthikumar. *Stochastic orders*. Springer Science & Business Media, 2007.

[76] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014. ISBN: 1107057132.

[77] Shai Shalev-Shwartz et al. "Learnability, stability and uniform convergence". In: *The Journal of Machine Learning Research* 11 (2010), pp. 2635–2670.

[78] Ohad Shamir. "On the complexity of bandit and derivative-free stochastic convex optimization". In: *Conference on Learning Theory*. 2013, pp. 3–24.

[79] Donald L Thistlethwaite and Donald T Campbell. "Regression-discontinuity analysis: An alternative to the ex post facto experiment." In: *Journal of Educational psychology* 51.6 (1960), p. 309.

[80] Twitter. *Twitter recommendation algorithm*. `https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm`.

[81] Twitter. *Twitter recommendation algorithm github repository*. `https://github.com/twitter/the-algorithm`.

[82] U.S Department of Education. *Issue brief: Early warning systems*. Available at `https://www2.ed.gov/rschstat/eval/high-school/early-warning-systems-brief.pdf`. Office of Planning, Evaluation and Policy Development, 2016. URL: `https://www2.ed.gov/rschstat/eval/high-school/early-warning-systems-brief.pdf`.

[83] U.S Senate Hearing. *No child left behind: The need to address the dropout crisis*. Available at `https://www.govinfo.gov/content/pkg/CHRG-107shrg83064/html/CHRG-107shrg83064.htm`. Committee on Health, Education, Labor, and Pensions, Nov. 2002. URL: `https://www.govinfo.gov/content/pkg/CHRG-107shrg83064/html/CHRG-107shrg83064.htm`.

[84] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.

[85] Cédric Villani. *Optimal transport: Old and new*. Vol. 338. Springer Science & Business Media, 2008.

[86] Cédric Villani. *Topics in optimal transportation*. 58. American Mathematical Society, 2003.

[87] Killian Wood, Gianluca Bianchin, and Emiliano Dall'Anese. "Online projected gradient descent for stochastic optimization with decision-dependent distributions". In: *IEEE Control Systems Letters* 6 (2021), pp. 1646–1651.

[88] Stephen J. Wright and Benjamin Recht. *Optimization for data analysis*. Cambridge University Press, 2022. DOI: `10.1017/9781009004282`.

[89] Shengjia Zhao et al. "Calibrating predictions to decisions: A novel approach to multi-class calibration". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 22313–22324.

[90] Yulai Zhao. "Optimizing the performative risk under weak convexity assumptions". In: *arXiv preprint arXiv:2209.00771* (2022).

[91] Tijana Zrnic et al. "Who leads and who follows in strategic classification?" In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 15257–15269.