Prediction and Statistical Inference in Feedback Loops

Tijana Zrnic

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2023-65
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2023/EECS-2023-65.html

May 5, 2023

Prediction and Statistical Inference in Feedback Loops

by

Tijana Zrnic

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Moritz Hardt, Co-chair
Professor Michael I. Jordan, Co-chair
Associate Professor William Fithian
Professor Bin Yu

Spring 2023

Prediction and Statistical Inference in Feedback Loops

Abstract

Prediction and Statistical Inference in Feedback Loops

by

Tijana Zrnic

Doctor of Philosophy in Electrical Engineering and Computer Sciences

University of California, Berkeley

Associate Professor Moritz Hardt, Co-chair

Professor Michael I. Jordan, Co-chair

Classical machine learning and statistics are built on the paradigm that there is a *fixed* quantity that we want to learn about a population, such as the best predictor of outcomes from features or the average effect of a treatment. In modern practices, however, predictions and inferences beget other predictions and inferences, causing the quantity of interest to change over time and drift away in a feedback loop. The feedback poses challenges for traditional methods, calling for new solutions. This thesis introduces new principles for prediction and inference in the presence of feedback loops.

The first part focuses on performative prediction. Performative prediction formalizes the phenomenon that predictive models—by means of being used to make consequential downstream decisions—often influence the outcomes they aim to predict in the first place. For example, travel time estimates on navigation apps influence traffic patterns and thus realized travel times, stock price predictions influence trading activity and hence prices. We examine common heuristics such as retraining, as well as more refined optimization strategies for dealing with performative feedback. At the end of the first part, we identify important scenarios where the act of prediction triggers feedback loops that are not explained by the framework of performativity, and we develop theory to describe and study such feedback.

The second part discusses principles for valid statistical inference, i.e., valid p-values and confidence intervals, in the presence of feedback. We consider two types of feedback: the first is due to data snooping, i.e., the practice of choosing which results to report only after seeing the data; the second arises when machine-learning systems are used to supply cheap predictions to augment or supplant high-quality data in future scientific analyses. In both cases, ignoring the feedback and naively applying classical statistical methods leads to inflated error rates and false discoveries; we provide alternative approaches that guarantee valid inferences in the face of feedback.

To my family, friends, teachers, mentors,
and everyone else who believed in me over the years.

Contents

Acknowledgments

This thesis is a summary of many years' worth of research conversations, group meeting discussions, hours spent staring at the whiteboard, long Zoom calls, and sleepless nights before deadlines. I was fortunate to share all of these with truly exceptional individuals. This thesis is as much a product of their efforts as it is mine.

I have to begin by thanking my advisors, Moritz and Mike, who played the biggest role in shaping my thesis work and the researcher I am today. I aspire to their academic breadth and unique and creative research styles, and I feel a deep sense of privilege to have been able to learn from them. They have been and continue to be role models of mine. Most importantly, they have made me feel endlessly supported. I am humbled by their belief in me from day one.

Next, I thank my thesis committee members, Will Fithian and Bin Yu. Will has been like a third advisor towards the end of my PhD. Working with him has strengthened my statistical mind and has been a great honor and pleasure. I am very grateful to Bin for all the feedback on my work, and I will remember many conversations with her fondly. I admire her commitment to both the finest technical details and big philosophical questions about the practice of data science.

I have been fortunate to have great mentors beyond my thesis committee. Vitaly Feldman hosted me for a summer internship at Apple, and working with him was the highlight of that year for me. Vitaly was an absolutely fantastic mentor. I have been lucky to share an office wall with Ben Recht, who has been a great conversationalist to have around Soda 5. Ben has done a great job keeping our office space lively and social. I am very thankful for his support in a few key moments of my PhD.

Every word in this thesis is a result of joint efforts with many wonderful collaborators. During my time at Berkeley I was lucky to collaborate with Anastasios Angelopoulos, Stephen Bates, Clara Fannjiang, Vitaly Feldman, Will Fithian, Paula Gradu, Moritz Hardt, Meena Jagadeesan, Daniel Jiang, Michael Jordan, Licong Lin, Eric Mazumdar, Celestine Mendler-Dünner, John Miller, Juan Perdomo, Aaditya Ramdas, Shankar Sastry, Martin Wainwright, and Yixin Wang. I am especially grateful to Aaditya for his guidance during the very first projects of my PhD. Beyond my collaborators, my research was greatly influenced by innumerable insightful discussions with members of SAIL, the Hardtcore group, and the EECS and Statistics Departments more generally; thank you for inspiring my work over the last six years.

Graduate school has been more fun and social than I ever could have imagined. To my friends—I don't want to double the page length of these acknowledgments so I won't include everyone by name, but you know who you are—thank you for making my time at Berkeley so special. I have many fond memories of fun dinners, drinks at Triple Rock and Jupiter, office banter, poker nights, Free Speech Fridays, beach and IM volleyball games. I even fondly remember several hikes. In addition to all the thank-yous, I want to express my sincerest apologies to everyone I yelled at during volleyball.

There are many people whose long-lasting support was instrumental in me starting my PhD at Berkeley in the first place. I am extremely grateful to Babak Hassibi, who was a wonderful mentor during a summer internship at Caltech back when I was an undergrad. He was the one who encouraged me to apply to PhD programs and this remains one of the best pieces of advice I ever got. I am also thankful to Dragana Bajovic for mentoring me through my undergraduate thesis. I learned so much from reading and discussing difficult technical material with her. Our interactions served as a great preparation for my PhD. Lastly, there are too many to list here, but I am indebted to all the great teachers and mentors I had throughout my education. They made me fall in love with learning and got me excited about math and engineering.

My deepest thank-yous go out to my family, who have been a constant source of encouragement despite all the miles between us, shaped every aspect of my personality, and taught me to be persistent and resilient. This thesis is a product of their unconditional support throughout my endeavors. Last but not least, I want to thank Anastasios for his love, patience, and kindness. I can't wait to see what life has in store for us.

Chapter 1

Introduction

Classical methods for data analysis were built on the assumption that there is one *fixed* learning target of interest, such as the best predictor of outcomes from features or the average effect of a treatment. Modern machine learning and statistics, however, are more dynamic and adaptive than ever before: predictions and inferences beget other predictions and inferences, causing *feedback loops* that shape and alter the target of learning. Such feedback loops invalidate the guarantees of classical learning methods, leading to inflated error rates and false discoveries. This thesis develops new principles for valid and reliable data analysis in the presence of feedback loops.

Feedback arises due to a number of factors, and what it means to successfully deal with feedback varies greatly depending on the context. We will focus on two broad goals: *prediction* and *statistical inference*. In prediction, our goal will be to solve a risk-minimization problem while taking into account its feedback-loop nature. We will introduce a framework called *performative prediction* for describing learning in feedback loops, and we will study a series of optimization strategies for learning under feedback. In statistical inference, our goal will be to compute valid p-values and confidence intervals, while allowing for the use of advanced computational methods in the data-analysis pipeline. Specifically, we will study *selective inference*—the practice of choosing or refining the inferential question of interest based on the data—as well as *prediction-powered inference*—the use of machine-learning predictions as data in downstream analyses. In the following paragraphs we describe the feedback-loop nature of each problem in more detail.

Predictions undoubtedly shape the world around us. One example setting where this is prominent is elections. Every high-profile election is accompanied by numerous high-profile forecasts of the election outcome. These forecasts have been observed to impact voter turnout [179], which in turns means that they impact the election outcome itself! Therefore, not only do predictions capture the patterns in our environment, but they also feed back into the environment and actively shape it. In the election example, a good forecasting mechanism should not only learn the preference over election candidates in the population at the time of polling, but it should predict the actual election outcome while taking into account the society's response responsibly and accurately.

Feedback also arises due to strategic incentives. There are numerous documented cases of individuals adapting strategically to algorithmic decision rules in order to achieve a desirable outcome, in domains ranging from social welfare programs to gig labor and social media moderation [24, 26, 128]. This widespread phenomenon, often known as Goodhart's law, can be summarized as: "When a measure becomes a target, it ceases to be a good measure" [156]. Due to their consequential nature, the algorithmic decisions cause people to alter their behavior and thus such decisions feed back into society. A natural target for the decision-maker is to learn a rule that yields high utility *after* the individuals have adapted to the decision rule, not merely on the data collected before the rule's deployment.

Another ubiquitous source of feedback in modern data analyses is data snooping, i.e., the practice of choosing which results to report based only *after* seeing the data. This practice, known in the literature as *selective inference*, offers more freedom to the analyst than the traditional paradigm of specifying the relevant hypotheses up front, but it also creates undesirable selection bias, thereby invalidating the error guarantees of classical statistical methods. The goal in selective inference is to compute valid p-values and confidence intervals, while allowing the analyst to adaptively refine the choice of statistical question in a data-driven manner.

Finally, feedback arises when predictions are leveraged as evidence in future scientific inquiry. Indeed, machine-learning algorithms are increasingly employed as black-box systems that supply predictions to augment or supplant costly experimental measurements. For example, accurate predictions of three-dimensional structures have been made for a vast catalog of known protein sequences [92, 165] and are now being used in proteomics studies [16]. Such predictions hold out the promise of increasing the pace and scope of scientific inquiry, however naively treating them as gold-standard data can naturally lead to false discoveries. We refer to the use of machine-learning predictions in downstream inferences as *prediction-powered inference*. The goal in prediction-powered inference is to compute valid p-values and confidence intervals, while making use of data sets imputed with machine-learning predictions to increase the effective sample size.

This thesis is based on works co-authored with Anastasios Angelopoulos, Stephen Bates, Clara Fannjiang, William Fithian, Moritz Hardt, Meena Jagadeesan, Michael I. Jordan, Eric Mazumdar, Celestine Mendler-Dünner, John Miller, Juan Perdomo, and S. Shankar Sastry [3, 75, 88, 123, 125, 137, 192, 193, 194].

Part I

Prediction in Feedback Loops

Chapter 2

Performative Prediction

Supervised learning excels at pattern recognition. When used to support consequential decisions, however, predictive models can trigger actions that influence the outcome they aim to predict. We call such predictions *performative*; the prediction causes a change in the distribution of the target variable.

Consider a simplified example of predicting credit default risk. A bank might estimate that a loan applicant has an elevated risk of default, and will act on it by assigning a high interest rate. In a self-fulfilling prophecy, the high interest rate further increases the customer's default risk. Put differently, the bank's predictive model is not calibrated to the outcomes that manifest from acting on the model.

Once recognized, performativity turns out to be ubiquitous. Traffic predictions influence traffic patterns, crime location prediction influences police allocations that may deter crime, recommendations shape preferences and thus consumption, stock price prediction determines trading activity and hence prices.

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. 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.

What would be desirable from the perspective of the decision-maker is an equilibrium where the model is optimal even after the data distribution reacts to its deployment. One such equilibrium notion coincides with the stable points of retraining; performativity therefore exposes retraining as a natural equilibrating dynamic rather than a nuisance.

This chapter formalizes *performative prediction*, tying together conceptual elements from statistical decision theory, causal reasoning, and game theory. The resulting framework raises many fundamental questions—for example, regarding the existence of stable points and other desirable equilibria, the behavior of retraining and other common optimization strategies—which will be addressed throughout the chapter.

The material in this chapter is based on works co-authored with Moritz Hardt, Meena Jagadeesan, Celestine Mendler-Dünner, John Miller, and Juan Perdomo [88, 123, 125, 137].

2.1 Framework

We put performativity at the center of 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 taking values in a closed, convex set $\Theta \subseteq \mathbb{R}^d$, that performs well on a fixed joint distribution $\mathcal{D}$ over covariates x and an outcome variable y.

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 a map $\mathcal{D}(\cdot)$—which we call the *distribution map*—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.

We now formally introduce the principal solution concepts of our framework: performative optimality and performative stability.

2.1.1 Performative optimality

In supervised learning, the goal is to learn a predictive model f_θ which minimizes the expected loss with respect to feature–outcome pairs (x, y) drawn i.i.d. from a fixed distribution $\mathcal{D}$. The optimal model $f_{\theta_{\text{SL}}}$ solves the following optimization problem,

$$\theta_{\text{SL}} = \operatorname*{arg\,min}_{\theta \in \Theta} \mathop{\mathbb{E}}_{z \sim \mathcal{D}} \ell(z; \theta),$$

where $\ell(z; \theta)$ denotes the loss of f_θ at a point z.

We contrast this with the *performative optimum*. As introduced previously, in settings where predictions support decisions, the manifested distribution over features and outcomes is in part determined by the deployed model. Instead of considering a fixed distribution $\mathcal{D}$, each model f_θ induces a potentially different distribution $\mathcal{D}(\theta)$ over instances z. A predictive model must therefore be evaluated with regard to the expected loss over the distribution $\mathcal{D}(\theta)$ it induces: its *performative risk*.

Definition 2.1.1 (Performative optimality and risk)**.** *A model $f_{\theta_{\text{PO}}}$ is* performatively optimal *if the following relationship holds:*

$$\theta_{\text{PO}} = \operatorname*{arg\,min}_{\theta \in \Theta} \mathop{\mathbb{E}}_{z \sim \mathcal{D}(\theta)} \ell(z; \theta).$$

We refer to $\text{PR}(\theta) \overset{\text{def}}{=} \mathbb{E}_{z \sim \mathcal{D}(\theta)} \ell(z; \theta)$ *as the* performative risk*; then,* $\theta_{\text{PO}} = \arg\min_{\theta \in \Theta} \text{PR}(\theta)$*.*

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

Example 2.1.1 (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 assigned score $f_\theta(x)$.*

In particular, define $\mathcal{D}(\theta)$ in the following way: x is a 1-dimensional feature supported on $\{\pm 1\}$ and $y \mid x \sim$ Bernoulli$(\frac{1}{2} + \mu x + \varepsilon\theta x)$ with $\mu \in (0, \frac{1}{2})$ and $\varepsilon < \frac{1}{2} - \mu$. 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: $\ell(z;\theta) = (y - f_\theta(x))^2$.

The parameter ε represents the performative aspect of the model. If $\varepsilon = 0$, 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 $f_{\theta_{\mathrm{SL}}}(x) = \mathbb{E}[y \mid x] = \frac{1}{2} + \mu x$, with $\theta_{\mathrm{SL}} = \mu$.

In the performative setting with $\varepsilon \neq 0$, 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

$$\theta_{\mathrm{PO}} = \operatorname*{arg\,min}_{\theta \in [0,1]} \mathop{\mathbb{E}}_{z \sim \mathcal{D}(\theta)} \left(y - \theta x - \frac{1}{2}\right)^2 \quad \iff \quad \theta_{\mathrm{PO}} = \frac{\mu}{1 - 2\varepsilon}.$$

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

2.1.2 Performative stability

A natural, desirable property of a model f_θ is that, given that we use the predictions of f_θ as a basis for decisions, those predictions are also simultaneously optimal for distribution that the model induces. We introduce the notion of *performative stability* to refer to predictive models that satisfy this property.

Definition 2.1.2 (Performative stability and decoupled risk)**.** *A model $f_{\theta_{\mathrm{PS}}}$ is* performatively stable *if the following relationship holds:*

$$\theta_{\mathrm{PS}} = \operatorname*{arg\,min}_{\theta \in \Theta} \mathop{\mathbb{E}}_{z \sim \mathcal{D}(\theta_{\mathrm{PS}})} \ell(z;\theta).$$

We refer to $\mathrm{DPR}(\theta, \theta') \overset{\text{def}}{=} \mathbb{E}_{z \sim \mathcal{D}(\theta)} \ell(z;\theta')$ *as the* decoupled performative risk*; then,* $\theta_{\mathrm{PS}} = \arg\min_{\theta \in \Theta} \mathrm{DPR}(\theta_{\mathrm{PS}}, \theta)$.

A performatively stable model $f_{\theta_{\mathrm{PS}}}$ minimizes the expected loss on the distribution $\mathcal{D}(\theta_{\mathrm{PS}})$ resulting from deploying $f_{\theta_{\mathrm{PS}}}$ in the first place. Therefore, a model that is performatively stable eliminates the need for retraining after deployment since any retraining procedure would simply return the same model parameters. Performatively stable models are *fixed points* of risk minimization. We further develop this idea in the next section.

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 2.1.2 (Example 2.1.1 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:*

$$\theta_{\text{PS}} = \operatorname*{arg\,min}_{\theta \in [0,1]} \mathop{\mathbb{E}}_{Z \sim \mathcal{D}(\theta_{\text{PS}})} \left(y - \theta x - \frac{1}{2} \right)^2 \quad \Longleftrightarrow \quad \theta_{\text{PS}} = \frac{\mu}{1 - \varepsilon}.$$

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 $\varepsilon = 0$. Note that, in general, if the 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 also refer to performative stability as simply stability.

Remark 2.1.1. *Notice that both performative stability and optimality can be expressed via the decoupled performative risk as follows:*

$$\begin{aligned} &\theta_{\text{PS}} \textit{ is performatively stable} && \Leftrightarrow && \theta_{\text{PS}} = \operatorname*{arg\,min}_{\theta} \ \mathrm{DPR}(\theta_{\text{PS}}, \theta), \\ &\theta_{\text{PO}} \textit{ is performatively optimal} && \Leftrightarrow && \theta_{\text{PO}} = \operatorname*{arg\,min}_{\theta} \ \mathrm{DPR}(\theta, \theta). \end{aligned}$$

2.1.3 Assumptions

It is easy to see that one cannot make any guarantees on the existence of stable points or algorithms for finding optima without making some regularity assumptions on $\mathcal{D}(\cdot)$. One reasonable way to quantify the regularity of $\mathcal{D}(\cdot)$ is to assume Lipschitz continuity; the Lipschitz constant determines how sensitive the induced distribution is to a change in model parameters. Intuitively, such an assumption captures the idea that, if decisions are made according to similar predictive models, then the resulting distributions over instances should also be similar. We now introduce this key assumption of our work, which we call *ε-sensitivity.*

Definition 2.1.3 (ε-sensitivity)**.** *We say that a distribution map $\mathcal{D}(\cdot)$ is* ε-sensitive *if for all $\theta, \theta' \in \Theta$:*

$$W_1\big(\mathcal{D}(\theta), \mathcal{D}(\theta')\big) \leqslant \varepsilon \|\theta - \theta'\|_2,$$

where W_1 denotes the Wasserstein-1 distance, or earth mover's distance.

The earth mover's distance is a natural notion of distance between probability distributions that provides access to a rich technical repertoire [169]. Furthermore, we can verify that it is satisfied in various settings.

Remark 2.1.2. *A simple example where this assumption is satisfied is for a Gaussian family. Given $\theta = (\mu, \sigma_1, \dots, \sigma_p) \in \mathbb{R}^{2p}$, define $\mathcal{D}(\theta) = \mathcal{N}(\varepsilon_1 \mu, \varepsilon_2^2 \, diag(\sigma_1^2, \dots, \sigma_p^2))$ where $\varepsilon_1, \varepsilon_2 \in \mathbb{R}$. Then $\mathcal{D}(\cdot)$ is ε-sensitive for $\varepsilon = \max\left\{|\varepsilon_1|, |\varepsilon_2|\right\}$.*

In addition to this assumption on the distribution map, we will often make standard assumptions on the loss function $\ell(z;\theta)$ which hold for broad classes of losses. To simplify our presentation, let $\mathcal{Z} \overset{\text{def}}{=} \cup_{\theta\in\Theta}\text{supp}(\mathcal{D}(\theta))$.

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

$$\left\|\nabla_\theta \ell(z;\theta) - \nabla_\theta \ell(z;\theta')\right\|_2 \leqslant \beta \left\|\theta - \theta'\right\|_2, \left\|\nabla_\theta \ell(z;\theta) - \nabla_\theta \ell(z';\theta)\right\|_2 \leqslant \beta \left\|z - z'\right\|_2, \tag{A1}$$

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

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

$$\ell(z;\theta) \geqslant \ell(z;\theta') + \nabla_\theta \ell(z;\theta')^\top(\theta - \theta') + \frac{\gamma}{2}\left\|\theta - \theta'\right\|_2^2, \tag{A2}$$

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

- (*second moment bound*) There exist constants σ^2 and L^2 such that for all $\theta, \theta' \in \Theta$:

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta)}\left[\|\nabla \ell(z;\theta')\|_2^2\right] \leqslant \sigma^2 + L^2\|\theta' - G(\theta)\|_2^2, \text{ where } G(\theta) \overset{\text{def}}{=} \arg\min_{\theta'} \mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta)} \ell(z;\theta'). \tag{A3}$$

The last assumption is customary in the stochastic optimization literature [17, 185].

2.2 Finding performatively stable points

Having introduced our framework for performative prediction, we now address some of the basic questions that arise in this setting and examine the behavior of common machine learning practices, such as retraining, through the lens of performativity.

As discussed previously, performatively stable models have the favorable property that they achieve minimal risk for the distribution they induce and hence eliminate the need for retraining. However, it is a priori not clear that such stable points exist; and even if they do exist, whether we can find them efficiently.

We begin to answer these questions by analyzing several different optimization strategies. The first is retraining, formally referred to as *repeated risk minimization* (RRM), where the exact minimizer is repeatedly computed on the distribution induced by the previous model parameters. We study RRM in Section 2.2.1. The second is *repeated gradient descent* (RGD), in which the model parameters are incrementally updated using a single gradient descent step on the objective defined by the previous iterate. We study RGD in Section 2.2.2. RGD is a computationally efficient approximation of RRM, which, as we show, adopts many favorable properties of RRM.

Both RRM and RGD are analyzed at the population-level, that is, assuming access to the *full* distribution $\mathcal{D}(\theta)$ when model θ is deployed. We relax this assumption in Section 2.2.3, where we study two variants of stochastic gradient descent called *greedy deploy* and *lazy deploy*, assuming only finite-sample access to $\mathcal{D}(\theta)$.

Our algorithmic analysis of these methods reveals the existence of stable points under the assumption that the distribution map $\mathcal{D}(\cdot)$ is sufficiently Lipschitz. We identify necessary and sufficient conditions for convergence to a performatively stable point.

2.2.1 Repeated risk minimization

We now formally define repeated risk minimization and prove one of our main results: sufficient and necessary conditions for retraining to converge to a performatively stable point.

Definition 2.2.1 (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 $t \geqslant 0$:*

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

Using a toy example, we again argue that restrictions on the map $\mathcal{D}(\cdot)$ are necessary to enable interesting analyses of RRM, otherwise it might be computationally infeasible to find performative optima, and performatively stable points might not even exist.

Example 2.2.1. *Consider optimizing the squared loss $\ell(z;\theta) = (y-\theta)^2$, where $\theta \in [0,1]$ and the distribution of the outcome y, according to $\mathcal{D}(\theta)$, is a point mass at 0 if $\theta \geqslant \frac{1}{2}$, and a point mass at 1 if $\theta < \frac{1}{2}$. Clearly there is no performatively stable point, and RRM will simply result in the alternating sequence* $1, 0, 1, 0, \ldots$. *The performative optimum in this case is $\theta_{\mathrm{PO}} = \frac{1}{2}$.*

To show convergence of retraining schemes, it is hence necessary to make a regularity assumption on $\mathcal{D}(\cdot)$, such as ε-sensitivity. We are now ready to state our main result regarding the convergence of repeated risk minimization.

Theorem 2.2.1. *Suppose that the loss $\ell(z;\theta)$ is β-jointly smooth* (A1) *and γ-strongly convex* (A2)*. If the distribution map $\mathcal{D}(\cdot)$ is ε-sensitive, then the following statements are true:*

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

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

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,

$$f(G(\theta)) - f(G(\theta')) \geqslant (G(\theta) - G(\theta'))^\top \nabla f(G(\theta')) + \frac{\gamma}{2}\|G(\theta) - G(\theta')\|_2^2 \tag{2.1}$$

$$f(G(\theta')) - f(G(\theta)) \geqslant \frac{\gamma}{2}\|G(\theta) - G(\theta')\|_2^2 \tag{2.2}$$

Together, these two inequalities imply that

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

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-jointly smooth. Using the dual formulation of the optimal transport distance (Lemma 2.5.1) and ε-sensitivity of $\mathcal{D}(\cdot)$,

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

Furthermore, using the first-order optimality conditions for convex functions, we have $(G(\theta) - G(\theta'))^\top \nabla f'(G(\theta')) \geqslant 0$, and hence $(G(\theta) - G(\theta'))^\top \nabla f(G(\theta')) \geqslant -\varepsilon\beta\|G(\theta) - G(\theta')\|_2\|\theta - \theta'\|_2$. Therefore, we conclude that,

$$-\gamma\|G(\theta) - G(\theta')\|_2^2 \geqslant -\varepsilon\beta\|G(\theta) - G(\theta')\|_2\|\theta - \theta'\|_2.$$

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 $G(\theta_{\mathrm{PS}}) = \theta_{\mathrm{PS}}$ by the definition of stability. Applying the result of part (a) yields

$$\|\theta_t - \theta_{\mathrm{PS}}\|_2 \leqslant \varepsilon\frac{\beta}{\gamma}\|\theta_{t-1} - \theta_{\mathrm{PS}}\|_2 \leqslant \left(\varepsilon\frac{\beta}{\gamma}\right)^t \|\theta_0 - \theta_{\mathrm{PS}}\|_2. \tag{2.3}$$

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

The main message of this theorem is that in performative prediction, if the loss function is sufficiently "nice" and the distribution map is sufficiently (in)sensitive, then one need only retrain a model a small number of times before it converges to a *unique* stable point.

One intriguing insight from our analysis is that this convergence result is in fact tight; removing any single assumption required for convergence by Theorem 2.2.1 is enough to construct a counterexample for which RRM diverges.

Proposition 2.2.1. *Suppose that the distribution map $\mathcal{D}(\cdot)$ is ε-sensitive with $\varepsilon > 0$. 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 and convex, but not strongly convex.

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

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

We provide counterexamples for all three statements in Section 2.5.

Proposition 2.2.1 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.

2.2.2 Repeated gradient descent

Theorem 2.2.1 demonstrates that repeated risk minimization converges to a unique performatively stable point if the sensitivity parameter ε is small enough. However, implementing RRM requires access to an exact optimization oracle. We now relax this requirement and demonstrate how a simple gradient descent algorithm also converges to a unique stable point.

Definition 2.2.2 (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 $t \geqslant 0$:*

$$\theta_{t+1} = G_{gd}(\theta_t) \overset{\text{def}}{=} \Pi_\Theta \left(\theta_t - \eta_t \mathop{\mathbb{E}}_{z \sim \mathcal{D}(\theta_t)} \nabla_\theta \ell(z; \theta_t) \right),$$

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

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. Like RRM, we can show that RGD is a contractive mapping for small enough sensitivity parameter ε.

Theorem 2.2.2. *Assume that the loss is β-jointly smooth* (A1) *and γ-strongly convex* (A2)*, and suppose that the distribution map $\mathcal{D}(\cdot)$ is ε-sensitive. Let $\varepsilon < \frac{\gamma}{\beta}$, and suppose that $\theta_{\text{PS}} \in$* Int(Θ) *. Then, repeated gradient descent (RGD) with a constant step size $\eta_t = \eta \overset{\text{def}}{=} \frac{\gamma - \varepsilon\beta}{2(1+\varepsilon^2)\beta^2}$ satisfies the following:*

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

(b) The iterates θ_t of RGD converge to the stable point θ_{PS} at a linear rate, $\|\theta_{t+1} - \theta_{\text{PS}}\|_2 \leqslant \delta$ for $t \geqslant \frac{2}{\eta(\gamma-\varepsilon\beta)} \log\left(\frac{\|\theta_1 - \theta_{\text{PS}}\|_2}{\delta}\right)$.

In fact, the upper bound $\varepsilon < \gamma/\beta$ on the sensitivity parameter is crucial for algorithmic convergence. It defines a regime outside which gradient descent is not guaranteed to converge to stability even at the population level.

Proposition 2.2.2. *Suppose that the distribution map $\mathcal{D}(\cdot)$ is ε-sensitive. Then, repeated gradient descent (RGD) can fail to converge to a performatively stable point in any of the following cases, for any choice of positive step size sequence $\{\eta_t\}_{t\geqslant 1}$:*

(a) The loss is β-jointly smooth (A1) *and convex, but not strongly convex* (A2)*, for any $\beta, \varepsilon > 0$.*

(b) The loss is β-jointly smooth (A1) *and γ-strongly convex* (A2)*, but $\varepsilon \geqslant \frac{\gamma}{\beta}$, for any $\gamma, \beta, \varepsilon > 0$.*

Therefore, γ/β is shown to be a sharp threshold for the convergence of gradient descent in performative settings, just like it was a sharp threshold for the convergence of repeated risk minimization.

2.2.3 Stochastic optimization

Repeated risk minimization and repeated gradient descent were defined as iterative algorithms that use the whole distribution $\mathcal{D}(\theta_t)$ for each update. In this section, we introduce two variants of stochastic gradient descent for optimization in performative settings with only finite samples. We refer to the two variants as *greedy deploy* and *lazy deploy*. Each method performs a stochastic gradient update using a single data point to the model parameters at every iteration, however they choose to deploy these updated models at different time intervals.

Greedy deploy

A natural algorithm for stochastic optimization in performative prediction is a direct extension of the stochastic gradient method, whereby at every time step, we observe a sample $z^{(k)} \sim \mathcal{D}(\theta_k)$, compute a gradient update to the current model parameters θ_k, and deploy the new model θ_{k+1} (see left panel in Figure 2.1). We call this algorithm *greedy deploy*.

While this procedure is algorithmically identical to the stochastic gradient method in traditional convex optimization, in performative prediction, the distribution of the observed samples depends on the trajectory of the algorithm. We begin by stating a technical lemma which introduces a recursion for the distance between θ_t and θ_{PS}.

Lemma 2.2.1. *Assume conditions* (A1), (A2), *and* (A3) *hold. If the distribution map $\mathcal{D}(\cdot)$ is ε-sensitive with $\varepsilon < \gamma/\beta$, then greedy deploy with step size η_t satisfies the following recursion for all $t \geqslant 1$:*

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

```
Greedy Deploy
Input: step size sequence $\{\eta_t\}_{t=1}^{\infty}$
Deploy initial classifier $\theta_1 \in \Theta$
For each $t = 1, 2, \dots$
    • Observe $z^{(t)} \sim \mathcal{D}(\theta_t)$
    • Update model parameters:
      $\theta_{t+1} = \Pi_\Theta(\theta_t - \eta_t \nabla \ell(z^{(t)}; \theta_t))$
    • Deploy $\theta_{t+1}$
```

```
Lazy Deploy
Input: step size sequence $\{\eta_{t,j}\}_{t,j=1}^{\infty}$
Deploy initial classifier $\theta_1 \in \Theta$
For each $t = 1, 2, \dots$
    • Set $\varphi_{t,1} = \theta_t$
    • For each $j = 1, \dots, n(t)$ :
        1. Observe $z_j^{(t)} \sim \mathcal{D}(\theta_t)$
        2. Update model parameters:
           $\varphi_{t,j+1} = \Pi_\Theta(\varphi_{t,j} - \eta_{t,j} \nabla \ell(z_j^{(t)}; \varphi_{t,j}))$
    • Deploy $\theta_{t+1} = \varphi_{t,n(t)+1}$
```

Figure 2.1: 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.

Similar recursions underlie many proofs of SGD, and Lemma 2.2.1 can be seen as their generalization to the performative setting. Furthermore, we see how the bound implies a strong contraction to the performatively stable point if the performative effects are weak, that is when $\varepsilon \ll \gamma/\beta$.

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 $O(1/t)$ rate.

Theorem 2.2.3. *Assume conditions* (A1), (A2), *and* (A3) *hold. If the distribution map $\mathcal{D}(\cdot)$ is ε-sensitive with $\varepsilon < \frac{\gamma}{\beta}$, then for all $t \geqslant 0$ greedy deploy with step size $\eta_t = \left((\gamma - \varepsilon\beta)t + 8L^2/(\gamma - \varepsilon\beta)\right)^{-1}$ satisfies*

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

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

Comparing this result to the traditional analysis of SGD for smooth, strongly convex objectives (e.g. [141]), we see that the traditional factor of γ is replaced by $\gamma - \varepsilon\beta$, which we view as the effective strong convexity parameter of the performative prediction problem. When $\varepsilon = 0$, 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.

This modification significantly changes the trajectory of lazy deploy relative to greedy deploy, given that the observed samples follow the distribution of the last *deployed* model, which might differ from the current iterate. More precisely, after deploying θ_t, we perform $n(t)$ stochastic gradient steps to the model parameters, using samples from $\mathcal{D}(\theta_t)$ before we deploy the last iterate as θ_{t+1} (see right panel in Figure 2.1).

At a high level, lazy deploy converges to performative stability because it progressively approximates repeated risk minimization (RRM). In Theorem 2.2.1 we showed that RRM converges to a performatively stable classifier at a linear rate when $\varepsilon < \gamma/\beta$. Since the underlying distribution remains static between deployments, a classical analysis of SGD shows that for large $n(t)$ these "offline" iterates $\varphi_{t,j}$ converge to the risk minimizer on the distribution corresponding to the previously deployed classifier. In particular, for large $n(t)$, $\theta_{t+1} \approx G(\theta_t)$. By virtue of approximately tracing out the trajectory of RRM, lazy deploy converges to θ_{PS} as well. This sketch is formalized in the following theorem.

Theorem 2.2.4. *Assume conditions* (A1), (A2), *and* (A3) *hold, and that the distribution map* $\mathcal{D}(\cdot)$ *is* ε*-sensitive with* $\varepsilon < \frac{\gamma}{\beta}$*. For any* $\alpha > 0$*, running lazy deploy with* $n(t) \geqslant n_0 t^\alpha$, $t = 1, 2, \ldots$ *many steps between deployments and step size sequence* $\eta_{t,j} = (\gamma j + 8L^2/\gamma)^{-1}$*, satisfies*

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

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

In this section we showed how varying the intervals at which we deploy models trained with stochastic gradient descent in performative settings leads to qualitatively different algorithms. While greedy deploy resembles classical SGD with a step size sequence adapted to the strength of distribution shift, lazy deploy can be viewed as a rough approximation of repeated risk minimization.

As we alluded to previously, the convergence behavior of both algorithms is critically affected by the strength of performative effects ε. For $\varepsilon \ll \gamma/\beta$, the effective strong convexity parameter $\gamma - \varepsilon\beta$ of the performative prediction problem is large. In this setting, the relevant distribution shift of deploying a new model is neglible 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 $\varphi_{t,j}$ to the risk minimizer on the current distribution $G(\theta_t)$ is unaffected by the value of ε. Lazy deploy then converges by closely mimicking the behavior of RRM.

Furthermore, both algorithms differ in their sensitivity to different initializations. In greedy deploy, the initial distance $\|\theta_1 - \theta_{\mathrm{PS}}\|_2^2$ decays polynomially, while in lazy deploy it decays at a linear rate. This suggests that the lazy deploy algorithm is more robust to poor initialization. While we derive these insights purely by inspecting our upper bounds, we find that these observations also hold empirically, as shown in the next section.

2.2.4 Experiments

We now proceed to illustrate the behavior of the discussed optimization strategies empirically, confirming our theoretical findings. As our main experimental setting, we use a strategic classification simulator available in the WhyNot package [124]. We begin by formally establishing how strategic classification can be cast as a performative prediction problem.

Stackelberg equilibria are performative optima

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

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:

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

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 [21, 76, 126], the distribution function $\mathcal{D}(\theta)$ in strategic classification corresponds to the data-generating process outlined in Figure 2.2.

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.

Setup

We run experiments on a dynamic credit scoring simulator in which an institution classifies the creditworthiness of loan applicants. As motivated previously, agents react to the insti-

```
**Input:** base distribution $\mathcal{D}$, classifier $f_\theta$, cost function $c$, and utility function $u$
**Sampling procedure for $\mathcal{D}(\theta)$:**
    1. Sample $(x, y) \sim \mathcal{D}$
    2. Compute best response $x_{\mathrm{BR}} \leftarrow \arg\max_{x'} u(x', \theta) - c(x', x)$
    3. Output sample $(x_{\mathrm{BR}},\ y)$
```

Figure 2.2: Distribution map for strategic classification.

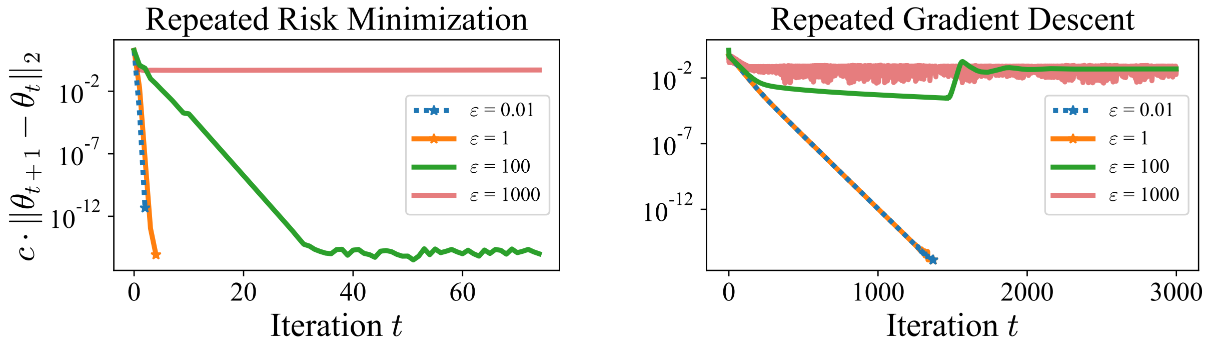


Figure 2.3: Convergence in domain of RRM (left) and RGD (right) for varying ε-sensitivity parameters. We add a marker if at the next iteration the distance between iterates is numerically zero. We normalize the distance by $c = \|\theta_{0,S}\|_2^{-1}$.

tution'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 2.2. For the base distribution $\mathcal{D}$, we use a class-balanced subset of a Kaggle credit scoring dataset [93]. 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 add a regularization term to the logistic loss to ensure that the objective is strongly convex. We assume that individuals have linear utilities $u(\theta, x) = -\langle\theta, x\rangle$ and quadratic costs $c(x', x) = \frac{1}{2\varepsilon}\|x' - x\|_2^2$, 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

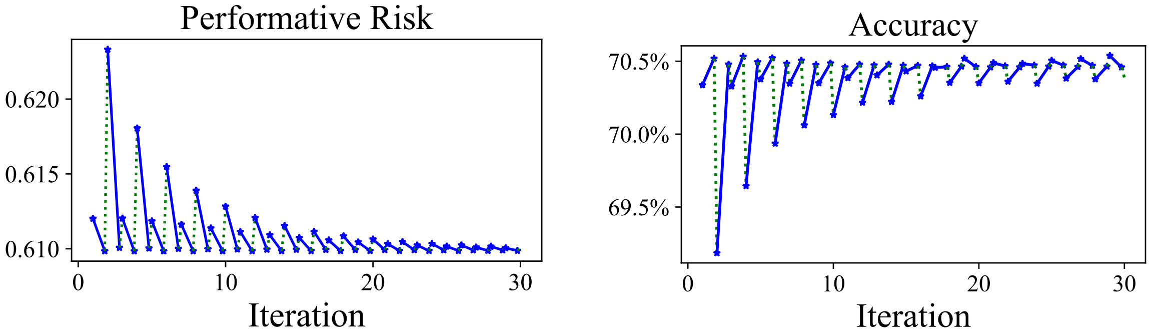


Figure 2.4: Performative risk (left) and accuracy (right) of the classifier θ_t at different stages of RRM for $\varepsilon = 80$. Blue lines indicates the optimization phase and green lines indicate the effect of the distribution shift after the classifier deployment.

open credit lines, and non-strategic features (e.g., age). Solving the optimization problem described in Figure 2.2, the best response for an individual corresponds to the following update,

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

where $x_S, x'_S, \theta_S \in \mathbb{R}^{|S|}$. As per convention in the literature [21, 76, 126], individual outcomes y are unaffected by strategic manipulation.

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 $x_S - \varepsilon\theta_S$ and to $x_S - \varepsilon\theta'_S$, respectively. The distance between these two shifted points is equal to $\varepsilon\|\theta_S - \theta'_S\|_2$. Since the optimal transport distance is bounded by $\varepsilon\|\theta - \theta'\|_2$ for every individual point, it is also bounded by this quantity over the entire distribution.

Repeated risk minimization and repeated gradient descent

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 2.3 (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 2.4. 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.

Next, we look at RGD. 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.

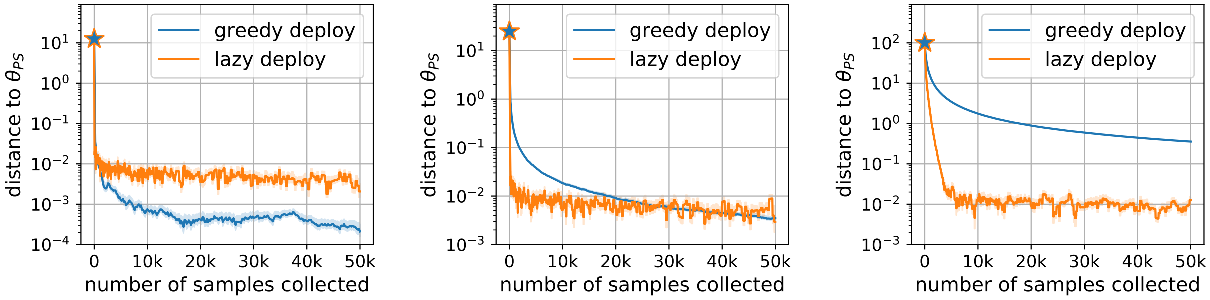


Figure 2.5: Convergence of lazy and greedy deploy to performative stability for varying values of $\varepsilon \in \{0.2, 0.6, 0.9\}$ (increased left to right). We use $n(t) = t$ for lazy deploy. The results are for the synthetic Gaussian example with $\mu = 10$, $\sigma = 0.1$.

Again, we can see in Figure 2.3 that the iterates converge for small values of ε and diverge for large values.

Greedy deploy and lazy deploy

We next empirically study greedy and lazy deploy. First, we carry out experiments using synthetic data where we can analytically compute stable points and carefully evaluate the tradeoffs suggested by our theory. Second, we evaluate the performance of these procedures on the same strategic classification simulator as in the previous section.

Synthetic data. For our first experiment, we consider the task of estimating the mean of a Gaussian random variable under performative effects. In particular, we consider minimizing the expected squared loss $\ell(z;\theta) = \frac{1}{2}(z-\theta)^2$ where $z \sim \mathcal{D}(\theta) = \mathcal{N}(\mu + \varepsilon\theta, \sigma^2)$. For $\varepsilon > 0$, the true mean of a distribution $\mathcal{D}(\theta)$ depends on our revealed estimate θ. Furthermore, for $\varepsilon < \gamma/\beta = 1$, the problem has a unique stable point. A short algebraic manipulation shows that $\theta_{\mathrm{PS}} = \frac{\mu}{1-\varepsilon}$. As per our theory, both greedy and lazy deploy converge to performative stability for all $\varepsilon < 1$.

We compare the convergence behavior of lazy deploy and greedy deploy for various values of ε in Figure 2.5. We choose step sizes for both algorithms according to our theorems in Section 2.2.3. In the case of lazy deploy, we set $\alpha = 1$, and hence $n(t) \propto t$. We see that when performative effects are weak, i.e. $\varepsilon \ll \gamma/\beta$, greedy deploy outperforms lazy deploy. Lazy deploy in turn is better at coping with large distribution shifts from strong performative effects. These results confirm the conclusions from our theory and show that the choice of whether to delay deployments or not can indeed have a large impact on algorithm performance depending on the value of ε.

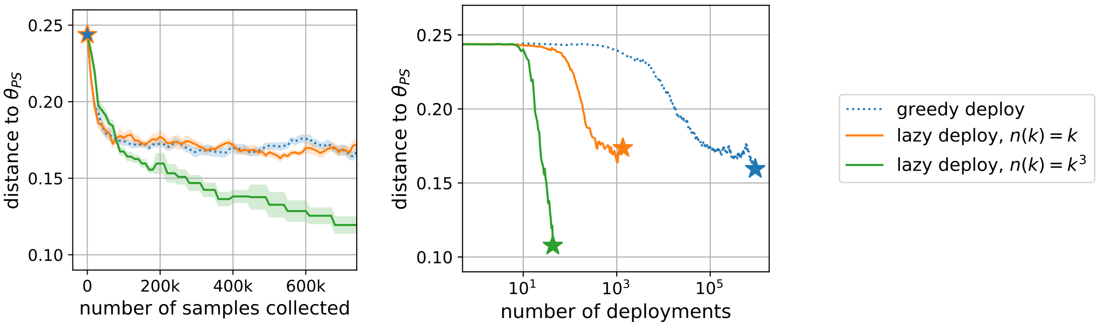


Figure 2.6: Convergence of lazy and greedy deploy to performative stability. Results are for the strategic classification experiments with $\varepsilon = 100$. (left panel) convergence as a function of the number of samples. (right panel) convergence as a function of the number of deployments.

Strategic classification. In addition to the experiments on synthetic data, we also evaluate the performance of the two optimization procedures in the previously described strategic classification setting. Since we now zoom in on stochastic optimization, 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. In contrast, in the previous section the learner observed 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 of this problem is much smaller than in the Gaussian example; we have $\gamma/\beta \approx 10^{-2}$. We explore the behavior of the algorithms outside the regime of provable convergence with $\varepsilon \gg \gamma/\beta$. We choose step sizes for both algorithms as defined in Section 2.2.3 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 2.2.3). As illustrated in Figure 2.6 (left), lazy significantly outperforms greedy deploy in this setting. Moreover, the performance of lazy deploy significantly improves with α. In addition to speeding up convergence, choosing larger sample collection schedules $n(t)$ substantially reduces the number of deployments, as seen in Figure 2.6 (right).

2.3 Finding performatively optimal points

So far we have discussed performative stability, which is a local definition of optimality by which a model minimizes the expected risk for the specific distribution that it induces. However, stability provides no general guarantees of performance beyond this equilibrium notion. In fact, stable models can have exceedingly poor performative risk, the central measure of the framework which captures the true risk incurred by the learner when deploying the model.

Reasoning by analogy, stable classifiers 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 classifiers minimize risk on the distribution that they induce, but they provide no global guarantees of performance.

Therefore, to develop accurate predictions in performative settings, we shift attention past performative stability and study optimizing the performative risk directly. This task will require a different algorithmic approach than the strategies analyzed previously, such as RRM and RGD. For instance, the learner needs to actively *anticipate* performative effects rather than myopically retrain until convergence, as the latter would only lead to stability.

First, in Section 2.3.1 we argue mathematically why performative stability is insufficient to guarantee satisfactory performance after model deployment. Then, we study strategies for finding performative optima in *convex* problems in Section 2.3.2. Finally, we study strategies for finding performative optima under great generally, even allowing *nonconvex* risk functions, in Section 2.3.3.

2.3.1 Contrasting optimality and stability

Up until now, all analyzed algorithmic strategies have been shown to converge to stable points. While the primary motivation for stability was eliminating the need for retraining, it may seem reasonable to think that stability ensures good performative risk as well. However, it turns out that there exist seemingly benign cases where the performative risk is strongly convex, but stable points actually *maximize* the performative risk.

Proposition 2.3.1. *For any $\gamma, \Delta > 0$, there exists a performative prediction problem where the loss is γ-strongly convex in θ, yet the unique stable point θ_{PS} maximizes the performative risk and $\mathrm{PR}(\theta_{\mathrm{PS}}) - \min_\theta \mathrm{PR}(\theta) \geqslant \Delta$.*

Proof. We prove the proposition by constructing an example. Let $z \sim \mathcal{D}(\theta)$ be a point mass at $\varepsilon\theta$, and define the loss to be:

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

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

$$\mathrm{PR}(\theta) = \left(\frac{\gamma}{2} - \varepsilon\beta\right) \cdot \|\theta\|_2^2. \tag{2.4}$$

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

In the above example, $\nabla_\theta \ell(z;\theta)$ is β-Lipschitz in z. The previous proposition thus shows that stable points can have an arbitrary suboptimality gap when $\varepsilon > \frac{\gamma}{2\beta}$. This is important since $\varepsilon < \frac{\gamma}{\beta}$ is the regime where the previously studied algorithms, namely repeated risk minimization and variants of gradient descent, converge to stability. Applying these methods when $\varepsilon \in (\gamma/(2\beta), \gamma/\beta)$ would hence maximize the performative risk on this problem.

We additionally point out that $\varepsilon = \frac{\gamma}{2\beta}$ is a sharp threshold for convexity of the performative risk in this example, as can be seen in equation (2.4). In the following section, we show that this threshold behavior is not an artifact of this particular setting, but rather a phenomenon that holds more generally.

2.3.2 Finding performatively optimal points under convexity

We begin by deriving key structural results illustrating how the performative risk can be convex in various natural 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 $\frac{\lambda}{2}\|\theta\|_2^2$ ensures $\mathrm{PR}(\theta)$ is convex.

In addition to the regularity conditions introduced at the beginning of the chapter, we will make repeated use of the following assumptions throughout the remainder of the section.

We will say that the loss is γ_z-strongly convex in z if for all $\theta \in \Theta$ and $z, z' \in \mathcal{Z}$,

$$\ell(z;\theta) \geqslant \ell(z';\theta) + \nabla_z \ell(z';\theta)^\top (z'-z) + \frac{\gamma_z}{2}\left\|z - z'\right\|_2^2. \tag{A4}$$

We state that 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 $\alpha \in (0,1)$:

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\alpha\theta+(1-\alpha)\theta')} \ell(z;\theta_0) \leqslant \mathop{\mathbb{E}}_{z\sim\alpha\mathcal{D}(\theta)+(1-\alpha)\mathcal{D}(\theta')} \ell(z;\theta_0). \tag{A5}$$

Smoothness and strong convexity conditions are standard in the optimization literature. 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*:

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

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 (A2) guarantees that DPR is convex in the second argument, mixture dominance (A5) 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.

On a more intuitive level, this assumption (A5) is essentially 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 [151]. Part of our analysis relies on incorporating tools from this literature, and we believe that further exploring technical connections between this field and performative prediction could be valuable. For example, using results from stochastic orders we can show that (A5) holds when the loss is convex in z and the distribution map $\mathcal{D}(\cdot)$ forms a *location-scale family* of the form:

$$z_\theta \sim \mathcal{D}(\theta) \;\Leftrightarrow\; z_\theta \overset{d}{=} (\Sigma_0 + \Sigma(\theta))z_0 + \mu_0 + \mu\theta, \tag{2.5}$$

where $z_0 \sim \mathcal{D}_0$ is a sample from a fixed zero-mean distribution $\mathcal{D}_0$, and $\Sigma(\theta), \mu$ are linear maps. Distribution maps of this sort are ubiquitous 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 simulator from Section 2.2 is a location family. 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 2.3.1. *Suppose that the loss function $\ell(z;\theta)$ is β-smooth in z* (A1)*, γ-strongly convex in θ* (A2)*, and that $\mathcal{D}(\cdot)$ is ε-sensitive. If mixture dominance* (A5) *holds, then the performative risk is λ-convex for $\lambda = \gamma - 2\varepsilon\beta$.*

Together with the example from the proof of Proposition 2.3.1, this theorem shows that $\frac{\gamma}{2\beta}$ is a sharp threshold for convexity of the performative risk. If ε is strictly less than this threshold, then under mixture dominance and appropriate conditions on the loss, the performative risk is strongly convex by Theorem 2.3.1. On the other hand, if ε is above this threshold, the example from Proposition 2.3.1 shows that there exists a performative prediction instance which satisfies the remaining assumptions, yet is non-convex; in particular, for $\varepsilon > \frac{\gamma}{2\beta}$ the performative risk is strictly concave in that example.

While the threshold $\varepsilon = \gamma/(2\beta)$ is in general tight, for certain families of distribution maps the conclusion of Theorem 2.3.1 can be made considerably stronger. Indeed, in some cases the performative risk is convex *regardless* of the magnitude of performative effects, as observed for the following location family.

Example 2.3.1. *Consider the following stylized model of predicting the final vote margin in an election contest. Features x, such as past polling averages, are drawn i.i.d. from a static distribution, $x \sim \mathcal{D}_x$. Since predicting a large margin in either direction can dissuade people from voting, we consider outcomes drawn from the conditional distribution: $y|x \sim g(x) + \mu^\top\theta + \xi$, where $g : \mathbb{R}^d \to \mathbb{R}$ is an arbitrary map, $\mu \in \mathbb{R}^d$ is a fixed vector, and ξ is a zero-mean noise variable. If ℓ is the squared loss, $\ell((x,y);\theta) = \frac{1}{2}(y - x^\top\theta)^2$, or the absolute loss, $\ell((x,y);\theta) = |y - x^\top\theta|$, then the performative risk is convex for any g and μ.*

The proof follows by simply observing that in both cases, the performative risk can be written as a linear function in θ composed with a convex function. Another interesting property of this example is that the distribution map is ε-sensitive with $\varepsilon = \|\mu\|_2$, yet the sensitivity parameter plays no role in the characterization of convexity. Motivated by this observation, we specialize the analysis in Theorem 2.3.1 to the particular case of location-scale families, and obtain a result that is at least as tight as the previous theorem.

Theorem 2.3.2. *Suppose that $\ell(z;\theta)$ is β-smooth* (A1)*,γ-strongly convex in θ* (A2)*, and γ_z-strongly convex in z* (A4)*. Furthermore, suppose that $\mathcal{D}(\theta)$ forms a location-scale family* (2.5) *with ε as its sensitivity parameter. Define Σ_{z_0} to be the covariance matrix of $z_0 \sim \mathcal{D}_0$, and let*

$$\sigma_{\min}(\mu) = \min_{\|\theta\|_2=1} \|\mu\theta\|_2, \sigma_{\min}(\Sigma) = \min_{\|\theta\|_2=1} \|\Sigma_{z_0}^{1/2}\Sigma(\theta)^\top\|_F.$$

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

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

This tighter bound leverages the fact that some losses are strongly convex in the performative variables, such as the squared loss when only the outcome variable exhibits performative effects. 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 for the sake of readability. We now illustrate an application of Theorem 2.3.2 with a scale family example.

Example 2.3.2. *Suppose that $x > 0$ is a one-dimensional feature drawn from a fixed distribution $\mathcal{D}_x$, and let $y|x \sim \theta x \cdot Exp(1)$ 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 2.3.2, the performative risk is convex and hence tractable to optimize, since $\gamma - \beta^2/\gamma_z = \mathbb{E}\,x^2 - (\mathbb{E}\,x)^2 \geqslant 0$ 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, whether stable points and performative optima now identify.

It turns out that both solutions can still have qualitatively different behavior, regardless of the strength of performative effects. First, notice that the example in the proof of Proposition 2.3.1 is a location family, and as such it satisfies mixture dominance. In that example, when $\varepsilon \in (\frac{\gamma}{2\beta}, \frac{\gamma}{\beta})$, 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 2.3.2, we provide another scale family example where the performative risk is convex regardless of ε, yet stable points can be arbitrarily suboptimal.

Example 2.3.3. *Suppose that $\mathcal{D}(\theta) = \mathcal{N}(\mu, \varepsilon^2\theta^2)$ for some $\mu \in \mathbb{R}$ and $\varepsilon > 0$. 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, $\theta_{\mathrm{PO}} = \mu/(1+\varepsilon^2)$.*

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 2.3.2, the performative risk is $\gamma - 2\varepsilon\beta + \gamma_z\sigma^2_{\min}(\Sigma) = 1 - 2\varepsilon + \varepsilon^2$-convex. Since $1 - 2\varepsilon + \varepsilon^2 = (\varepsilon - 1)^2 \geqslant 0$, the performative risk is always convex and hence tractable to optimize.

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 classifier deployments. Without knowledge of the underlying distribution map, it is infeasible to 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 $z \sim \mathcal{N}(\varepsilon\theta, 1)$ be a one-dimensional Gaussian and let $\ell(z;\theta) = \frac{1}{2}(z-\theta)^2$ be the squared loss. Then, the performative risk, $\mathrm{PR}(\theta) = \frac{1}{2}(\varepsilon-1)^2\theta^2$, is a simple, convex function for all values of ε (as indeed confirmed by Theorem 2.3.2, since $\gamma - 2\varepsilon\beta + \gamma_z\sigma^2_{\min}(\mu) = 1 - 2\varepsilon + \varepsilon^2 \geqslant 0$). However, gradients are unavailable since they depend on the density of $\mathcal{D}(\theta)$, denoted p_θ, which is typically unknown:

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

Despite the simplicity of this example, earlier approaches to optimization such as RRM 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, RRM computes the sequence of updates $\theta_{t+1} = \arg\min_\theta \mathbb{E}_{z\sim\mathcal{D}(\theta_t)} \frac{1}{2}(z-\theta)^2 = \varepsilon\theta_t$, which diverges for $|\varepsilon| > 1$.

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 [1, 62, 153]. 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. [62] sample a vector $u \sim \mathrm{Unif}(\mathcal{S}^{d-1})$ to get a slightly biased gradient estimator,

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

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. However, their rate of convergence can be slow and scales poorly with the problem dimension. In general, zeroth-order methods require $\widetilde{O}(d^2/\Delta^2)$ samples to obtain a Δ-suboptimal point [1, 153], 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:

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

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

Algorithm 1 Two-Stage Algorithm for Location Families

```
Stage 1: Construct a model of the distribution map
// Estimate location parameter $\mu$ with experiment design
for $i = 1$ to $n$ do
    -Sample and deploy classifier $\theta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$.
    -Observe $z_i \sim \mathcal{D}(\theta_i)$.
end for
-Estimate $\mu$ via ordinary 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 $j = n + 1$ to $2n$ do
    -Deploy classifier $\theta_j = 0$, and observe $z_j \sim \mathcal{D}(0)$.
end for
Stage 2: Minimize a finite-sample approximation of the performative risk, $\arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{j=n+1}^{2n} \ell(z_j + \hat{\mu}\theta; \theta)$.
```

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

$$z_\theta \sim \mathcal{D}(\theta) \;\Leftrightarrow\; z_\theta \overset{d}{=} z_0 + \mu\theta,$$

where the matrix $\mu \in \mathbb{R}^{m\times d}$ is an unknown parameter, and $z_0 \sim \mathcal{D}_0$ is a zero-mean random variable.[1]

As discussed previously, location-scale families encompass many formal examples discussed in prior work. They capture the intuition that in performative settings, the data points are composed of a *base* component z_0, representing the natural data distribution in the absence of performativity, and an additive performative term.

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 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_0, without performative effects. With both of these components, given any θ', we can simulate $z \sim \widehat{\mathcal{D}}(\theta')$ by taking $z = z_0 + \hat{\mu}\theta'$.

[1] 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 1 is made for convenience. In general, any subgaussian distribution with full rank covariance would suffice.

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

$$\widehat{\mathrm{PR}}(\theta) = \mathop{\mathbb{E}}_{z\sim\widehat{\mathcal{D}}(\theta)} \ell(z;\theta) = \mathop{\mathbb{E}}_{z_0\sim\mathcal{D}_0} \ell(z_0 + \hat{\mu}\theta;\theta).$$

Note that $\mathrm{PR}(\theta) = \mathbb{E}_{z_0\sim\mathcal{D}_0} \ell(z_0 + \mu\theta;\theta)$. 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:

$$\hat{\theta}_n \in \mathop{\arg\min}_{\theta\in\Theta} \widehat{\mathrm{PR}}_n(\theta) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=n+1}^{2n} \ell(z_i + \hat{\mu}\theta;\theta).$$

The main technical result in this section shows that, under appropriate regularity assumptions on the loss, Algorithm 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.

Theorem 2.3.3 (Informal)**.** *Under appropriate smoothness and strong convexity assumptions on the loss ℓ, if the distribution of z_0 is subgaussian, and if the number of samples $n \geqslant \Omega\left(d + m + \log(1/\delta)\right)$, then, with probability $1-\delta$, Algorithm 1 returns a point $\hat{\theta}_n$ such that*

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

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.

2.3.3 Finding performatively optimal points without convexity

Finally, we consider optimization in the face performativity under pretty great generality, allowing arbitrary, possibly *nonconvex* losses and making no structural assumptions on the distribution map $\mathcal{D}(\theta)$. Due to the inherent uncertainty about $\mathcal{D}(\theta)$, it is not possible to find a model with low performative risk offline. The learner needs to interact with the environment and deploy models θ to explore the induced distributions $\mathcal{D}(\theta)$. Given the online nature of this task, we measure the loss incurred by deploying a sequence of models $\theta_1, \ldots, \theta_T$ by evaluating the *performative regret*:

$$\mathrm{Reg}(T) := \sum_{t=1}^{T} \Big(\mathbb{E}\ \mathrm{PR}(\theta_t) - \min_{\theta} \mathrm{PR}(\theta)\Big),$$

where the expectation is taken over the possible randomness in the choice of $\{\theta_t\}_{t=1}^T$. Performative regret measures the suboptimality of the deployed sequence of models relative to a performative optimum $\theta_{\mathrm{PO}} \in \arg\min_\theta \mathrm{PR}(\theta)$.

At first glance, performative regret minimization might seem equivalent to a classical bandit problem. Bandit solutions minimize regret while requiring only noisy zeroth-order access to the unknown reward function—in our case PR. The resulting regret bounds generally grow with some notion of complexity of the reward function.

However, a naive application of bandit baselines misses out on a crucial fact: performative regret minimization exhibits significantly richer feedback than bandit feedback. When deploying a model θ, the learner gains access to samples from the induced distribution $\mathcal{D}(\theta)$, rather than only a noisy estimate of the risk $\mathrm{PR}(\theta)$. We call this feedback model *performative feedback*. Together with the fact that the learner knows the loss $\ell(z;\theta)$, performative feedback can be used to inform the reward of unexplored arms. For instance, it allows the computation of an unbiased estimate of $\mathbb{E}_{z\sim\mathcal{D}(\theta)}\, \ell(z;\theta')$ for *any* point θ'.

To illustrate the power of this feedback model, consider the limiting case in which the performative effects entirely vanish and the distribution map is constant, i.e. $\mathcal{D}(\theta) \equiv \mathcal{D}_*$ for some fixed distribution $\mathcal{D}_*$ independent of θ. With zeroth-order feedback, the learner would still need to deploy different models to explore the landscape of PR and find a point with low risk. However, with performative feedback, a *single* deployment gives samples from $\mathcal{D}_*$, thus resolving all uncertainty in the landscape of PR apart from finite-sample uncertainty. This raises the question: *with performative feedback, can one achieve regret bounds that scale only with the complexity of the distribution map, and not that of the performative risk?*

Set up as an online learning problem, performative prediction can be formalized as follows. At every time step t, the learner chooses a model θ_t and observes a constant number m_0 of i.i.d. samples,

$$\{z_t^{(i)}\}_{i\in[m_0]}, \text{ where } z_t^{(i)} \sim \mathcal{D}(\theta_t).$$

The regret incurred by choosing θ_t at time step t is $\Delta(\theta_t) := \mathrm{PR}(\theta_t) - \mathrm{PR}(\theta_{\mathrm{PO}})$, where θ_{PO} is the performative optimum. For simplicity, we assume $\max\{\|\theta\| : \theta \in \Theta\} \leqslant 1$ and $\ell(z;\theta) \in [0,1]$ for all z and θ.

The constant m_0 quantifies how many samples the learner can collect in a time window determined by how often they incur regret. For example, at the beginning of each week the learner might update the model, and thus at the end of each week they incur regret for the model they chose to deploy. In that case, m_0 is the number of samples the learner collects per week. Note that a learner with larger m_0 collects an empirical distribution that more accurately reflects $\mathcal{D}(\theta_t)$ and thus naturally minimizes regret at a faster rate.

A black-box bandits approach

Performative regret minimization can be set up as a continuum-armed bandits problem where an arm corresponds to a choice of model parameters θ. Performative feedback is sufficient to simulate noisy zeroth-order feedback about the reward function, as assumed in bandits.

When we deploy θ_t, the samples from $\mathcal{D}(\theta_t)$ enable us to compute an unbiased estimate

$$\widehat{\mathrm{PR}}(\theta_t) = \frac{1}{m_0}\sum_{i=1}^{m_0} \ell\big(z_t^{(i)};\theta_t\big)$$

of the risk $\mathrm{PR}(\theta_t)$. Moreover, since we assume the loss function is bounded, the noise in the estimate $\widehat{\mathrm{PR}}(\theta_t)$ is subgaussian, as typically required in bandits.

A standard condition that makes continuum-armed bandit problems tractable is a bound on how fast the reward can change when moving from one arm to a nearby arm. Formally, this regularity is ensured by assuming Lipschitzness of the reward function—in our case, Lipschitzness of the performative risk.

The dependence of $\mathrm{PR}(\theta)$ on θ is twofold: through the loss argument and through the distribution argument. Thus, the most natural way to ensure that $\mathrm{PR}(\theta)$ is Lipschitz is to ensure that each of these two dependencies is Lipschitz. This yields the following bound:

Lemma 2.3.1 (Lipschitzness of PR). *If the loss $\ell(z;\theta)$ is L_z-Lipschitz in z and L_θ-Lipschitz in θ and the distribution map is ε-sensitive, then the performative risk is $(L_\theta+\varepsilon L_z)$-Lipschitz.*

The intuition behind Lemma 2.3.1 is that $\mathrm{PR}(\theta)$ is guaranteed to be Lipschitz if $\mathrm{DPR}(\theta,\theta')$ is Lipschitz in each argument individually. Lipschitzness in the second argument follows from requiring that the loss be Lipschitz in θ. Lipschitzness in the first argument follows from combining Lipschitzness of the loss in z and ε-sensitivity of the distribution map.

Once we have established Lipschitzness of the performative risk, we can apply techniques from the Lipschitz bandits literature. Kleinberg et al. [100] proposed a bandit algorithm that adaptively discretizes promising regions of the space of arms, using Lipschitzness of the reward function to bound the additional loss due to discretization. Their method, called the *zooming algorithm*, will serve as a baseline for our problem. The algorithm enjoys an instance-dependent regret that takes advantage of nice problem instances, while maintaining tight guarantees in the worst case. The rate depends on the *zooming dimension*, which is upper bounded in the worst case by the dimension of the full space d.

Proposition 2.3.2 (Zooming algorithm [100]). *Suppose $m_0 = o(\log T)$. Then, after T deployments, the zooming algorithm achieves a regret bound of*

$$\mathrm{Reg}(T) = \mathcal{O}\left(T^{\frac{d_0+1}{d_0+2}}\left(\frac{\log T}{m_0}\right)^{\frac{1}{d_0+2}}(L_\theta+\varepsilon L_z)^{\frac{d_0}{d_0+2}}\right),$$

where d_0 denotes the $(L_\theta+\varepsilon L_z)$-zooming dimension.

The zooming dimension quantifies the niceness of a problem instance by measuring the size of a covering of near-optimal arms, instead of the entire parameter space. Roughly speaking, if the reward function is very "flat" in that there are many near-optimal points,

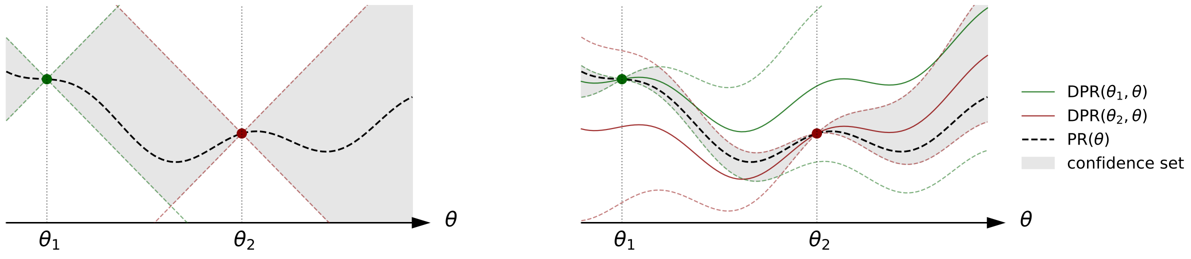


Figure 2.7: Confidence bounds after deploying θ_1 and θ_2. (left) Confidence bounds via Lipschitzness, as stated in equation (2.6). (right) Performative confidence bounds, as stated in equation (2.7).

then the zooming dimension is close to the dimension d of the parameter space. However, if the reward has sufficient curvature, then the zooming dimension can be much smaller than d. The zooming dimension is defined formally as follows:

Definition 2.3.1 (α-zooming dimension)**.** *A performative prediction problem instance has α-zooming dimension equal to d_0 if any minimal s-cover of any subset of $\{\theta : \Delta(\theta) \leqslant 16\alpha s\}$ includes at most a constant multiple of $(3/s)^{d_0}$ elements from $\{\theta : 16\alpha r \leqslant \Delta(\theta) < 32\alpha r\}$, for all $0 < r \leqslant s \leqslant 1$.*

For well-behaved instances, the definition intuitively requires every minimal s-cover of $\{\theta : 16\alpha r \leqslant \Delta(\theta) < 32\alpha r\}$ to have size at most of order $(3/s)^{d_0}$. Definition 2.3.1 slightly differs from the definition presented in [100] and makes the dependence on the Lipschitz constant explicit; we use Definition 2.3.1 to later ease the comparison to our new algorithm. The differences between the two definitions are minor technicalities that we do not expect to alter the zooming dimension in a meaningful way, neither formally nor conceptually.

Making use of performative feedback

We now illustrate how we can take advantage of performative feedback beyond computing a point estimate of the deployed model's risk. For now, we ignore finite-sample considerations and assume access to the entire distribution $\mathcal{D}(\theta)$ after deploying a model θ. We will address finite-sample uncertainty when presenting our main algorithm in the next section.

First, we demonstrate how performative feedback allows constructing tighter confidence bounds on the performative risk of unexplored models, compared to only relying on Lipschitzness of the risk function $\mathrm{PR}(\theta)$.

Suppose we deploy a set of models $\mathcal{S} \subseteq \Theta$ and for each $\theta \in \mathcal{S}$ we observe $\mathcal{D}(\theta)$. Then, under the regularity conditions of Lemma 2.3.1, we can bound the risk of any $\theta' \in \Theta$ as

$$\max_{\theta \in \mathcal{S}} \mathrm{PR}(\theta) - (L_\theta + L_z\varepsilon)\|\theta - \theta'\| \leqslant \mathrm{PR}(\theta') \leqslant \min_{\theta \in \mathcal{S}} \mathrm{PR}(\theta) + (L_\theta + L_z\varepsilon)\|\theta - \theta'\|. \tag{2.6}$$

These confidence bounds only use $\mathcal{D}(\theta)$ for the purpose of computing $\mathrm{PR}(\theta)$ and rely on Lipschitzness to construct confidence sets around the risk of unexplored models. However, in light of the structure of PR, the bounds in equation (2.6) do not make full use of performative feedback; in particular, access to $\mathcal{D}(\theta)$ actually allows us to evaluate $\mathrm{DPR}(\theta, \theta')$ for *any* θ'. Importantly, this information can further reduce our uncertainty about $\mathrm{PR}(\theta')$, and we can bound:

$$\begin{aligned}\mathrm{PR}(\theta') &= \mathrm{DPR}(\theta, \theta') + (\mathrm{DPR}(\theta', \theta') - \mathrm{DPR}(\theta, \theta')) \\ &\leqslant \mathrm{DPR}(\theta, \theta') + L_z\varepsilon\|\theta - \theta'\|.\end{aligned}$$

Thus we can get tighter bounds on the performative risk at an unexplored parameter θ':

$$\max_{\theta\in\mathcal{S}} \mathrm{DPR}(\theta, \theta') - L_z\varepsilon\|\theta - \theta'\| \leqslant \mathrm{PR}(\theta') \leqslant \min_{\theta\in\mathcal{S}} \mathrm{DPR}(\theta, \theta') + L_z\varepsilon\|\theta - \theta'\|. \tag{2.7}$$

We call the confidence bounds computed in (2.7) *performative confidence bounds*. In Figure 2.7, we visualize and contrast these confidence bounds with the confidence bounds obtained via Lipschitzness. We observe that by computing DPR we can significantly tighten the confidence regions.

The tightness of the confidence bounds depends on the set $\mathcal{S}$ of deployed models. By choosing a cover of the parameter space, we can get an estimate of the performative risk that has low approximation error on the whole parameter space.

Proposition 2.3.3. *Let $\mathcal{S}_\gamma$ be a γ-cover of Θ and suppose we deploy all models $\theta \in \mathcal{S}_\gamma$. Then, using performative feedback we can compute an estimate of the performative risk $\widehat{\mathrm{PR}}(\theta)$ such that for any $\theta \in \Theta$ it holds that*

$$|\mathrm{PR}(\theta) - \widehat{\mathrm{PR}}(\theta)| \leqslant \gamma L_z\varepsilon.$$

Proposition 2.3.3 implies that after exploring the cover $\mathcal{S}_\gamma$, we can find a model whose suboptimality is at most $\mathcal{O}(\gamma L_z\varepsilon)$. To contextualize the bound in Proposition 2.3.3, consider an approach that uses the same cover $\mathcal{S}_\gamma$ but only relies on zeroth-order feedback, that is, $\{\mathrm{PR}(\theta) : \theta \in \mathcal{S}_\gamma\}$. Then, the only feasible estimate of PR over the whole space is $\widehat{\mathrm{PR}}(\theta) = \mathrm{PR}(\Pi_{\mathcal{S}_\gamma}(\theta))$, where $\Pi_{\mathcal{S}_\gamma}(\theta) = \arg\min_{\theta'\in\mathcal{S}_\gamma} \|\theta - \theta'\|$ is the projection onto the cover $\mathcal{S}_\gamma$. This zeroth-order approach only guarantees an accuracy of $|\mathrm{PR}(\theta) - \widehat{\mathrm{PR}}(\theta)| \leqslant (L_z\varepsilon + L_\theta)\gamma$, a strictly weaker approximation than the one in Proposition 2.3.3.

Sequential elimination of suboptimal models

Now we show how performative confidence bounds can guide exploration. Specifically, we show that every deployment informs the risk of unexplored models, which allows us to sequentially discard suboptimal regions of the parameter space.

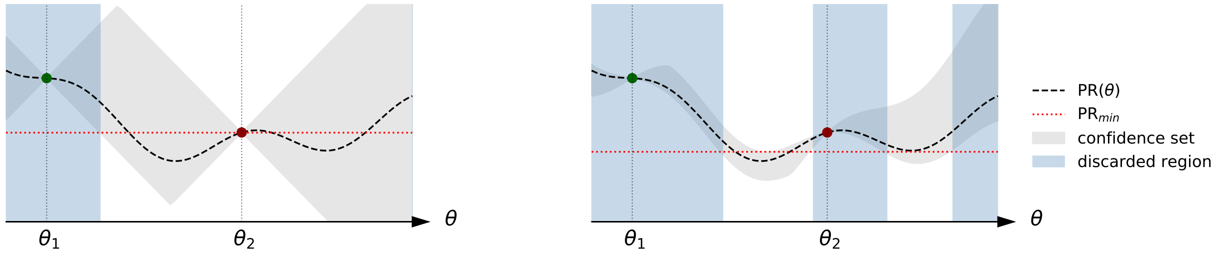


Figure 2.8: (left) Baseline confidence bounds. (right) Performative confidence bounds. Performative feedback allows discarding unexplored suboptimal models even in regions that have not been explored. A model θ is discarded if $\mathrm{PR}_{\mathrm{LB}}(\theta) > \mathrm{PR}_{\min}$. The loss function and feedback model are the same as in Figure 2.7.

To develop a formal procedure for discarding points, let $\mathrm{PR}_{\mathrm{LB}}(\theta)$ denote a lower confidence bound on $\mathrm{PR}(\theta)$ and $\mathrm{PR}_{\min}$ denote an upper confidence bound on $\mathrm{PR}(\theta_{\mathrm{PO}})$ based on the information from the models deployed so far:

$$\begin{aligned}\mathrm{PR}_{\mathrm{LB}}(\theta) &= \max_{\theta' \text{ already deployed}} \left(\mathrm{DPR}(\theta', \theta) - L_z\varepsilon\|\theta - \theta'\|\right),\\ \mathrm{PR}_{\min} &= \min_{\theta\in\Theta} \min_{\theta' \text{ already deployed}} \left(\mathrm{DPR}(\theta', \theta) + L_z\varepsilon\|\theta' - \theta\|\right).\end{aligned}$$

It is not difficult to see that the following lower bound on the suboptimality of model θ holds:

Proposition 2.3.4. *For all $\theta \in \Theta$, we have $\Delta(\theta) \geqslant \mathrm{PR}_{\mathrm{LB}}(\theta) - \mathrm{PR}_{\min}$.*

In particular, models θ with $\mathrm{PR}_{\mathrm{LB}}(\theta) > \mathrm{PR}_{\min}$ cannot be optimal. We recall our toy example from Figure 2.7 and illustrate in Figure 2.8 the parameter configurations we can discard after the deployment of two models, θ_1 and θ_2. We can see that access to DPR allows us to discard a large portion of the parameter space, and, in contrast to the baseline black-box approach, it is possible to discard regions of the space that have not been explored.

Performative confidence bounds algorithm

We introduce our main algorithm that builds on the two insights from the previous section. We furthermore provide a rigorous, finite-sample analysis of its guarantees. Our *performative confidence bounds* algorithm, formally stated in Algorithm 2, takes advantage of performative feedback by assessing the risk of unexplored models and thus guiding exploration. We give an overview of the main steps.

Inspired by the successive elimination algorithm [59], the algorithm keeps track of and refines an *active* set of models $\mathcal{A} \subseteq \Theta$. Roughly speaking, active models are those that are estimated to have low risk and only they are admissible to deploy. To deal with finite-sample uncertainty, the algorithm proceeds in phases which progressively refine the precision of the finite-sample risk estimates. More precisely, in phase p the algorithm chooses an error tolerance γ_p and deploys a model for n_p steps. In each step m_0 samples induced by the deployed model are collected, and n_p is chosen so that the inferred estimates of DPR are γ_p-accurate. Formally, if θ is deployed in phase p, we collect an empirical distribution $\widehat{\mathcal{D}}(\theta)$ of $n_p m_0$ samples so that $|\widehat{\mathrm{DPR}}(\theta,\theta') - \mathrm{DPR}(\theta,\theta')| \leqslant \gamma_p$ for all θ' with high probability, where

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

These estimates of DPR are used to construct performative confidence bounds and refine $\mathcal{A}$.

Each phase begins by constructing a net of the current active set $\mathcal{A}$. The points in the net are sequentially deployed in the phase, unless they are deemed to be suboptimal based on previous deployments in that phase and are in that case eliminated. During phase p, we denote by $\mathcal{P}_p$ the running set of deployed points and by $\mathcal{S}_p$ the running set of net points that have not been discarded. We initialize $\mathcal{S}_p$ to a minimal r_p-net of the current set of active points $\mathcal{A}$, denoted $\mathcal{N}_{r_p}(\mathcal{A})$, where r_p is proportional to γ_p. A net point θ gets eliminated from $\mathcal{S}_p$ if no point in $\mathrm{Ball}_{r_p}(\theta) := \{\theta' \in \Theta : \|\theta' - \theta\| \leqslant r_p\}$ is active. This means that we may deploy suboptimal points in the net if they help inform active points nearby.

Before we state the regret bound for Algorithm 2, let us comment on an important component in the analysis. Recall that throughout the algorithm we operate with finite-sample estimates of the decoupled performative risk to bound the risk of unexplored models. Specifically, for any deployed θ, we make use of $\widehat{\mathrm{DPR}}(\theta,\theta')$ for *all* θ'. Since we need these estimates to be valid simultaneously for all θ', we rely on uniform convergence. As such, the Rademacher complexity of the loss function class naturally enters the bound.

Definition 2.3.2 (Rademacher complexity). *Given a loss function $\ell(z;\theta)$, we define $\mathfrak{C}^*(\ell)$ to be:*

$$\mathfrak{C}^*(\ell) = \sup_{\theta\in\Theta}\ \sup_{n\in\mathbb{N}}\ \sqrt{n}\cdot \mathop{\mathbb{E}}_{\varepsilon, z^\theta}\left(\sup_{\theta'\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j \ell(z_j^\theta;\theta')\right|\right),$$

where $\varepsilon_j \sim$ Rademacher and $z_j^\theta \sim \mathcal{D}(\theta)$, $\forall j \in [n]$, which are all independent of each other.

Now we can state our regret guarantee for Algorithm 2.

Theorem 2.3.4. *Assume the loss $\ell(z;\theta)$ is L_z-Lipschitz in z and let ε denote the sensitivity of the distribution map. Suppose that $\mathfrak{C}$ is any value such that $\mathfrak{C}^*(\ell) \leqslant \mathfrak{C}$ and $m_0 = o(\mathcal{B}^2_{\log T,\mathfrak{C}})$, where $\mathcal{B}_{\log T,\mathfrak{C}} := \sqrt{\log T} + \mathfrak{C}$. Then, after T time steps, Algorithm 2 achieves a regret bound of*

$$\mathrm{Reg}(T) = \mathcal{O}\left(T^{\frac{d_0+1}{d_0+2}}\left(\frac{(L_z\varepsilon)^{d_0}\mathcal{B}^2_{\log T,\mathfrak{C}}}{m_0}\right)^{\frac{1}{d_0+2}} + \sqrt{T}\,\frac{\mathcal{B}_{\log T,\mathfrak{C}}}{\sqrt{m_0}}\right),$$

where d_0 is the $(L_z\varepsilon)$-sequential zooming dimension (see Definition 2.3.3).

```
Algorithm 2 Performative Confidence Bounds Algorithm
Require: time horizon $T$, number of samples collected per step $m_0$, sensitivity parameter $\varepsilon$, Lipschitz constant $L_z$, complexity bound $\mathfrak{C}$
  Initialize $\mathcal{A} \leftarrow \Theta$
  for phase $p = 0, 1, \ldots$ do
      Set error tolerance $\gamma_p = 2^{-p}$ and net radius $r_p = \frac{\gamma_p}{L_z \varepsilon}$
      Let $n_p = \left\lceil \frac{\left(2\mathfrak{C}+3\sqrt{\log T}\right)^2}{\gamma_p^2 m_0} \right\rceil$
      Initialize $\mathcal{S}_p \leftarrow \mathcal{N}_{r_p}(\mathcal{A})$    // Initialize $\mathcal{S}_p$ to minimal $r_p$-cover of $\mathcal{A}$
      Initialize $\mathcal{P}_p \leftarrow \emptyset$
    while $\mathcal{S}_p \neq \emptyset$ do
          Draw $\theta_{\mathrm{net}} \in \mathcal{S}_p$ uniformly at random
          Deploy $\theta_t$ for $n_p$ steps to form $\widehat{\mathrm{DPR}}(\theta_{\mathrm{net}}, \cdot)$
          $\mathcal{S}_p \leftarrow \mathcal{S}_p \setminus \theta_{\mathrm{net}}$
          $\mathcal{P}_p \leftarrow \mathcal{P}_p \cup \theta_{\mathrm{net}}$ // Update set of deployed models
          $\mathrm{PR}_{\min} \leftarrow \min_{\theta\in\Theta} \min_{\theta'\in\mathcal{P}_p} \widehat{\mathrm{DPR}}(\theta', \theta) + L_z\varepsilon\|\theta' - \theta\|$ // Update estimate of $\mathrm{PR}(\theta_{\mathrm{PO}})$
          $\mathrm{PR}_{\mathrm{LB}}(\theta) \leftarrow \max_{\theta'\in\mathcal{P}_p} \left(\widehat{\mathrm{DPR}}(\theta', \theta) - L_z\varepsilon\|\theta' - \theta\|\right), \forall\theta \in \mathcal{A}$ // Update LB for models
          $\mathcal{A} \leftarrow \mathcal{A} \setminus \{\theta \in \mathcal{A} : \mathrm{PR}_{\mathrm{LB}}(\theta) > \mathrm{PR}_{\min} + 2\gamma_p\}$ // Update active region
          $\mathcal{S}_p \leftarrow \mathcal{S}_p \setminus \{\theta \in \mathcal{S}_p : \mathrm{Ball}_{r_p}(\theta) \cap \mathcal{A} = \emptyset\}$ // Remove deactivated net points
    end while
  end for
```

Remark 2.3.1 (Consequences for finding performative optima)**.** *Algorithm 2 has the additional property that it generates a model with near-minimal performative risk. In particular, an intermediate step in the proof of Theorem 2.3.4 shows if T is sufficiently large, the final iterate θ_T of Algorithm 2 satisfies:*

$$\mathbb{E}\left[\mathrm{PR}(\theta_T) - \min_{\theta\in\Theta}\mathrm{PR}(\theta)\right] \leqslant \mathcal{O}\left(T^{-\frac{1}{d_0+2}}\left(\frac{(L_z\varepsilon)^{d_0}\mathcal{B}^2_{\log T,\mathfrak{C}}}{m_0}\right)^{\frac{1}{d_0+2}}\right),$$

where d_0 is the $(L_z\varepsilon)$-zooming dimension.

Notice that the regret in Theorem 2.3.4 depends on the sequential zooming dimension (formally defined in Definition 2.3.3). This sequential variant of zooming dimension accounts for the sequential elimination of models within each phase. We will show in the next section that the sequential zooming dimension is upper bounded by the usual zooming dimension (see Proposition 2.3.5).

The primary advantage of Theorem 2.3.4 over the Lipschitz bandit baseline can be seen by examining the first term in the regret bound. This term resembles the black-box regret bound from Proposition 2.3.2; however, the key difference is that that the bound of Theorem 2.3.4

depends on the complexity of the distribution map rather than that of the performative risk. In particular, the Lipschitz constant is $L_z\varepsilon$ and not $L_\theta + L_z\varepsilon$. The advantage is pronounced when $\varepsilon \to 0$, making the first term of the bound in Theorem 2.3.4 vanish so only the $\mathcal{O}(\sqrt{T})$ term remains. On the other hand, the bound in Proposition 2.3.5 maintains an exponential dimension dependence.

Taking the limit as $\varepsilon \to 0$ also reveals why the second term in the bound emerges. Even if the distribution map is constant, there is regret arising from finite-sample error. This is a key conceptual difference in the meaning of Lipschitzness of the distribution map versus that of the performative risk: $L_\theta + L_z\varepsilon$ being 0 implies that PR is flat and thus all models are optimal, while performative regret minimization is nontrivial even if $L_z\varepsilon = 0$. Unlike the first term, the second term due to finite samples is dimension-independent apart from any dependence implicit in the Rademacher complexity.

We note that the presence of the Rademacher complexity term $\mathfrak{C}^*(\ell)$ makes a direct comparison of the bound in Theorem 2.3.4 and the bound in Proposition 2.3.5 subtle. When the Rademacher complexity is very high, the regret bound in Theorem 2.3.4 may be worse. Nonetheless, for many natural function classes, the Rademacher complexity is polynomial in the dimension; in these cases, Theorem 2.3.4 can substantially outperform the regret bound in Proposition 2.3.5.

Another key feature of the regret bound in Theorem 2.3.4 worth highlighting is the zooming dimension. Definition 2.3.1 allows us to directly compare the dimension in Theorem 2.3.4 with the dimension in Proposition 2.3.5: the $(L_z\varepsilon)$-zooming dimension of Algorithm 2 is no larger than, and most likely smaller than, the $(L_\theta + L_z\varepsilon)$-zooming dimension in the black-box approach. Moreover, the sequential variant of zooming dimension in Theorem 2.3.4 can further reduce the dimension.

Finally, the main assumption underpinning the bound in Theorem 2.3.4 is that DPR is $(L_z\varepsilon)$-Lipschitz in its first argument. Sensitivity (Def. 2.1.3) coupled with Lipschitzness of the loss in the data achieves this. However, this property can hold with different regularity assumptions on the distribution map and loss function; e.g., if the loss is bounded and the distribution map is Lipschitz in total variation distance.

The zooming dimension of Definition 2.3.1 does not take into account that, using performative feedback, our algorithm can eliminate unexplored models *within* a phase. We illustrate the benefits of this sequential exploration strategy in Figure 2.9, where the deployment of two models is sufficient to eliminate the remaining model in the cover. This motivates a sequential definition of zooming dimension that captures the benefits of sequential exploration.

To set up the definition of *sequential zooming dimension*, we need to introduce some notation. For a set of points $\mathcal{S}$, enumeration $\pi : \mathcal{S} \to \{1, \dots, |\mathcal{S}|\}$ that specifies an ordering

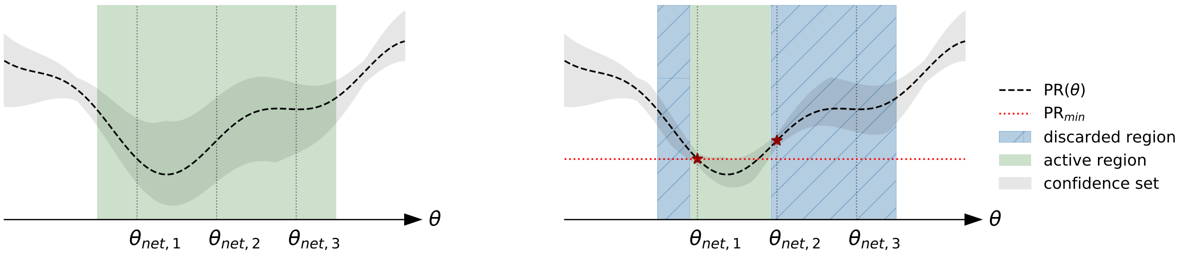


Figure 2.9: Sequential deployment of models allows Algorithm 2 to eliminate points from $\mathcal{S}_p$, reducing the number of deployments during the phase. We see how the deployment of $\theta_{\text{net},1}$ and $\theta_{\text{net},2}$ allows one to eliminate $\theta_{\text{net},3}$.

on $\mathcal{S}$, and number $k \in \{1, \dots, |\mathcal{S}|\}$, let

$$\begin{aligned}
\text{PR}_{\text{LB}}(\theta; k) &:= \max_{\theta' \in \mathcal{S}: \pi(\theta') < k} \left(\text{DPR}(\theta', \theta) - L_z \varepsilon \|\theta - \theta'\|\right), \\
\text{PR}^s_{\text{LB}}(k) &:= \min_{\theta \in \text{Ball}_s(\pi^{-1}(k))} \text{PR}_{\text{LB}}(\theta; k), \\
\text{PR}_{\min}(k) &:= \min_{\theta} \min_{\theta' \in \mathcal{S}: \pi(\theta') < k} \left(\text{DPR}(\theta', \theta) + L_z \varepsilon \|\theta' - \theta\|\right).
\end{aligned}$$

Here, $\text{PR}_{\text{LB}}(\theta; k)$ is a lower bound on $\text{PR}(\theta)$ arising from the first $k-1$ deployments of the phase. Similarly, $\text{PR}^s_{\text{LB}}(k)$ captures the minimal lower confidence bound on the performative risk for any point in an s-ball around the k-th deployed model, $\pi^{-1}(k)$. Finally, $\text{PR}_{\min}(k)$ captures an upper bound on $\text{PR}(\theta_{\text{PO}})$, estimated from the first $k-1$ deployments.

Using the above terms, we see that $\text{PR}^s_{\text{LB}}(k) \leqslant \text{PR}_{\min}(k) + 4\alpha s$ is the population version of the condition that a model in the cover does not get discarded. The sequential zooming dimension captures the maximal number of models in each suboptimality band that can be deployed.

Definition 2.3.3 (Sequential zooming dimension)**.** *A performative prediction problem instance has α-zooming dimension equal to d_0 if for any minimal s-cover $\mathcal{S}$ of any subset of $\{\theta : \Delta(\theta) \leqslant 16\alpha s\}$ and all $0 < r \leqslant s \leqslant 1$, the expected number of models $\theta \in \mathcal{S} \cap \{\theta : 16\alpha r \leqslant \Delta(\theta) < 32\alpha r\}$ with*

$$\text{PR}^s_{\text{LB}}(\pi(\theta)) \leqslant \text{PR}_{\min}(\pi(\theta)) + 4\alpha s \tag{2.8}$$

is at most a constant multiple of $(3/s)^{d_0}$, where the expectation is taken over a uniformly sampled enumeration $\pi : \mathcal{S} \to \{1, \dots, |\mathcal{S}|\}$.

The sequential zooming dimension is bounded by the zooming dimension in Definition 2.3.1.

Proposition 2.3.5. *For all $\alpha > 0$, the α-zooming dimension is at least as large as the α-sequential zooming dimension.*

The claim of Proposition 2.3.5 follows by definition. To see this, let d_0 be the α-zooming dimension. This means that $\mathcal{S}$ includes at most a constant multiple of $(3/s)^{d_0}$ elements from $\{\theta : 16\alpha r \leqslant \Delta(\theta) < 32\alpha r\}$, for all $0 < r \leqslant s \leqslant 1$. This immediately guarantees that the subset of $\mathcal{S}$ characterized by (2.8) is at most a multiple of $(3/s)^{d_0}$, as desired.

Regret minimization for location families

We show how further knowledge about the structure of the distribution map can help reduce the complexity of performative regret minimization, without necessarily implying favorable structure of the performative risk. Once again, we apply our guiding principle of focusing exploration on learning the distribution map. Since the loss function is known, we can extrapolate knowledge about the distribution map to estimate the performative risk.

We focus on the previously introduced setting of location families, which are distribution maps that depend on θ via a linear shift. More precisely, location families are distribution maps of the form $z \sim \mathcal{D}(\theta) \Leftrightarrow z \overset{d}{=} z_0 + \mu_*^\top \theta$, where $\mu_* \in \mathbb{R}^{d_\Theta \times m}$ is an unknown matrix and $z_0 \in \mathbb{R}^m$ is a zero-mean subgaussian sample from a base distribution $\mathcal{D}_0$.

At a high level, our algorithm can be described as follows: at every step t, the learner deploys a model θ_t and collects m_0 samples from $\mathcal{D}(\theta_t)$. We will write $\bar{z}_t := \frac{1}{m_0}\sum_{i=1}^{m_0} z_t^{(i)}$ for the corresponding sample average at time t. Then, based on all samples collected so far, the algorithm computes the least-squares estimate of μ_* along with a confidence region for μ_*. In the next step the algorithm picks the model that minimizes a lower confidence bound $\mathrm{PR}_{\mathrm{LB}}(\theta)$. See Algorithm 3 for details.

Algorithm 3 Performative Regret Minimization for Location Families

```
Require: time horizon $T$, number of samples collected per step $m_0$, base distribution $\mathcal{D}_0$,
  bound $M_*$ such that $\|\mu_*\| \leqslant M_*$
  Initialize confidence set $\mathcal{C}_1 \leftarrow \{\mu : \|\mu\| \leqslant M_*\}$
  for step $t = 1, 2, \ldots$ do
      $\mathrm{PR}_{\mathrm{LB}}(\theta) \leftarrow \min_{\mu \in \mathcal{C}_t} \mathbb{E}_{z_0 \sim \mathcal{D}_0} \ell(z_0 + \mu\theta; \theta) \; \forall \theta \in \Theta$ // Update LB for all models
      Deploy $\theta_t = \arg\min_\theta \mathrm{PR}_{\mathrm{LB}}(\theta)$ // Deploy model with lowest LB
      Compute $\bar{z}_t = \frac{1}{m_0}\sum_{i=1}^{m_0} z_t^{(i)}$ from collected samples
      Let $\Sigma_t \leftarrow \sum_{i=1}^{t} \theta_i \theta_i^\top + \frac{1}{m_0} I$
      $\hat{\mu}_t \leftarrow \Sigma_t^{-1}\left(\sum_{i=1}^{t} \theta_i \bar{z}_i^\top\right)$ // Update estimate of $\mu_*$
      $\mathcal{C}_{t+1} \leftarrow \left\{\mu : \left\|\Sigma_t^{1/2}(\hat{\mu}_t - \mu)\right\| < \frac{M_* + \sqrt{8m_0 + 8\log T + 2d_\Theta \log\left(1 + \frac{Tm_0}{d_\Theta}\right)}}{\sqrt{m_0}}\right\}$ // Update conf. set
  end for
```

This algorithm is inspired by LinUCB [109], a standard bandits algorithm for linear rewards whose regret scales as $\tilde{\mathcal{O}}(d\sqrt{T})$, where d is the dimension of the linear map. Importantly, unlike in the LinUCB analysis, our objective function $\mathrm{PR}(\theta)$ is *not* linear in θ. Still, the nature of performative feedback allows us to learn the hidden linear structure in the distribution map and apply this knowledge to obtain confidence bounds on the performative risk. Below we state our algorithm for performative regret minimization for location families together with its regret guarantees.

Theorem 2.3.5. *Suppose that $\ell(z;\theta)$ is L_z-Lipschitz in z, $\mathcal{D}_0$ is 1-subgaussian, and $m_0 = o(\log T)$. Then, after T time steps, Algorithm 3 achieves a regret bound of*

$$\mathrm{Reg}(T) = \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{m_0}}\max\{L_z, 1\}\sqrt{T}\max\left\{d, \sqrt{dm}\right\}\right).$$

Remark 2.3.2. *For simplicity, we assume that $\mathcal{D}_0$ is known in Algorithm 3. This assumption is justified, for example, when we have plenty of historical data about a population, before any model deployment. We note that Theorem 2.3.5 can be extended to the case where we only have a finite data set from $\mathcal{D}_0$, by relying on a uniform convergence argument.*

Theorem 2.3.5 shows that by leveraging the hidden linear structure of the distribution map, Algorithm 3 inherits the $\tilde{\mathcal{O}}(\sqrt{T})$ rate of LinUCB. This bears resemblance to the regret bound in Theorem 2.3.4 that also scaled primarily with the complexity of the distribution map. Furthermore, similarly to Algorithm 2, we see that the regret bound for Algorithm 3 holds while allowing the loss to have arbitrary dependence on θ. For example, the loss need not be convex and, as a result, the performative risk need not be convex either.

We conclude by comparing Theorem 2.3.5 to Theorem 2.3.3, which provided an algorithm for finding performative optima for location families in the special case when the performative risk is *strongly convex*. Converting the previous optimization error into a regret bound yields a bound of $\mathcal{O}(\sqrt{T}(d+m))$. While this bears resemblance to Theorem 2.3.5, the rates are not directly comparable. Algorithm 1 does not assume knowledge of the base distribution $\mathcal{D}_0$, but rather deploys the model $\theta = 0$ in initial steps to collect samples from $\mathcal{D}_0$ (see Remark 2.3.2 for how to combine this strategy with our algorithm). In any case, the main benefit of Theorem 2.3.5 is that it applies to a more general setting, placing significantly fewer restrictions on the loss function and the performative risk.

2.3.4 Experiments

We complement our theoretical findings with an empirical evaluation of different methods on two tasks: the strategic classification simulator from Section 2.2 and a synthetic linear regression example.

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 from Section 2.3.2, versus optimization algorithms for finding stable

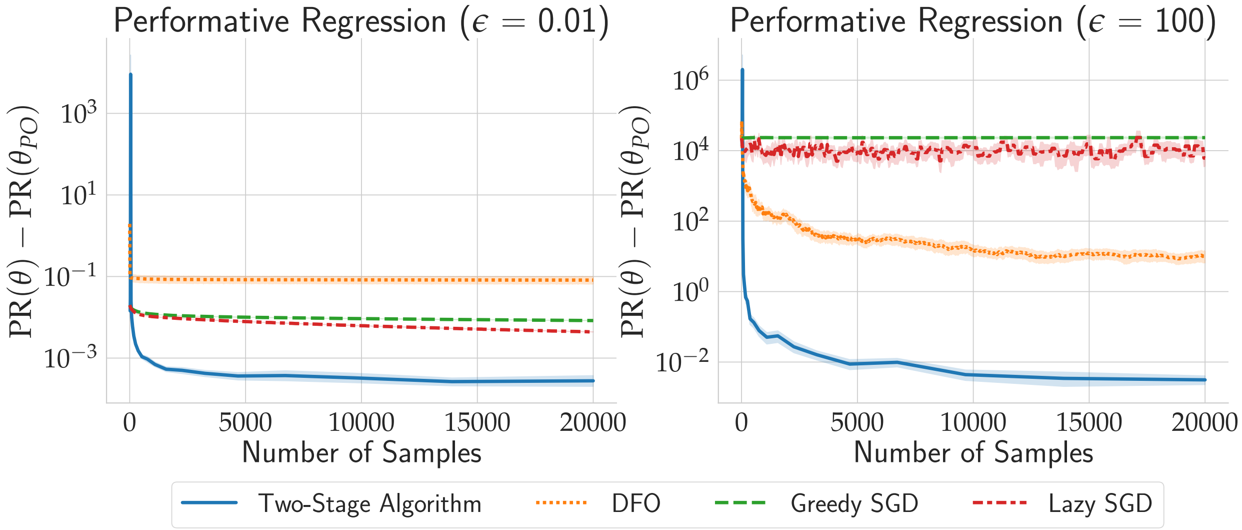


Figure 2.10: Suboptimality gap versus number of samples collected for the two-stage algorithm, DFO algorithm, greedy deploy, and lazy deploy, for $\varepsilon = 0.01$ (left) and $\varepsilon = 100$ (right). Each experiment is repeated 50 times; we display 95% bootstrap confidence intervals.

points, in particular greedy deploy and lazy deploy. In addition, we focus on highlighting the differences in the sample efficiency of the different algorithms and examine their sensitivity to the relevant structural assumptions outlined in Section 2.3.2. To evaluate derivative-free methods, we implement the "gradient descent without a gradient" algorithm from [62], 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.

Linear regression experiments. We begin by evaluating how increasing the strength of performative effects affects the behavior of the different optimization procedures in settings where the performative risk is convex. We recall the setup from Example 2.3.1, where the learner attempts to solve a linear regression with performative labels. Given a parameter θ, data are drawn from $\mathcal{D}(\theta)$ according to:

$$x \sim \mathcal{N}(0, \Sigma_x),\ U_y \sim \mathcal{N}(0, \sigma_y^2),\ y = \beta^\top x + \mu^\top \theta + U_y.$$

This distribution map is a location family, and is ε-sensitive with $\varepsilon = \|\mu\|_2$. Performance is measured according to the squared loss, $\ell((x,y);\theta) = \frac{1}{2}(y - \theta^\top x)^2$. Furthermore, the performative risk is convex for all choices of μ.

For small ε, we see that greedy and lazy SGD converge to a stable point that approximately minimizes the performative risk (see left panel in Figure 2.10). However, as we

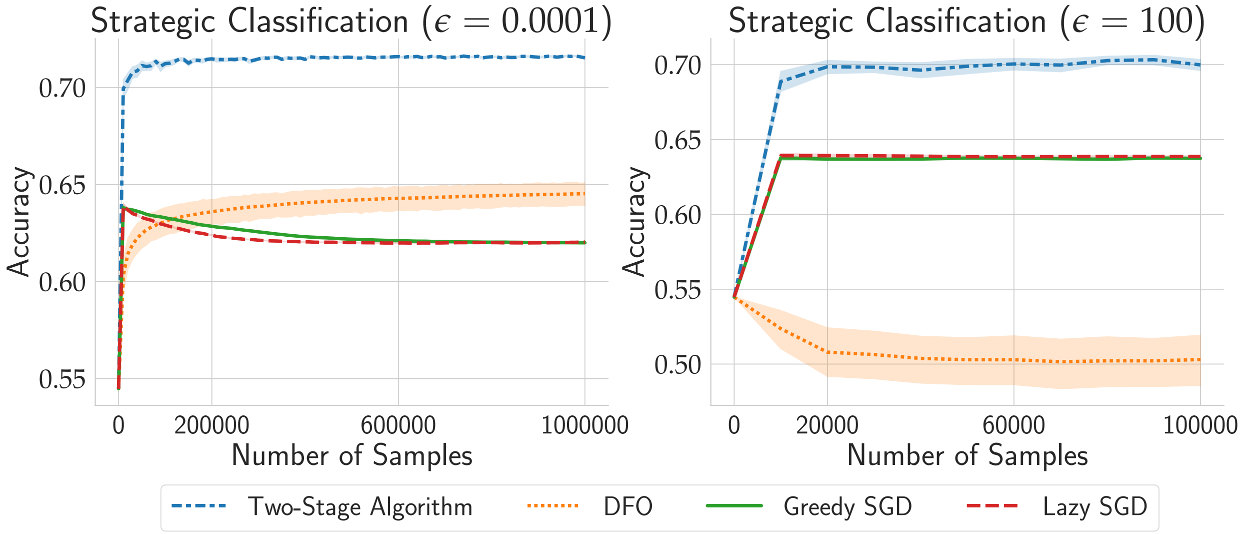


Figure 2.11: Classification accuracy versus number of samples collected for the two-stage algorithm, DFO algorithm, greedy deploy, and lazy deploy, for $\varepsilon = 0.0001 \leqslant \frac{\gamma}{2\beta}$ (left) and $\varepsilon = 100 \gg \frac{\gamma}{2\beta}$ (right). Each experiment is repeated 50 times; we display 95% bootstrap confidence intervals.

increase the strength of performative effects, these methods fail to make any progress, and are outperformed by both the DFO algorithm and the two-stage approach by a considerable margin (see right panel in Figure 2.10). The two-stage procedure efficiently converges after a small number of samples and its behavior is largely unaffected as we increase the value of ε, while the DFO algorithm becomes considerably slower when ε is large.

Strategic classification. We next consider experiments on the credit scoring simulator from Section 2.2. Since the logistic loss is not strongly convex in the features, we only have a certificate of convexity when ε is small enough (namely, $\varepsilon \leqslant \frac{\gamma}{2\beta}$). We consider two values of ε: one which is below this critical threshold, and one large value for which we do not have theoretical guarantees. 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 2.11). Together with the linear regression experiments, this observation serves as further evidence that stable points have significantly worse performative risk relative to performative optima, even in regimes where $\varepsilon < \gamma/(2\beta)$. Note also 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 $\varepsilon \gg \gamma/(2\beta)$ which is outside the regime of our theoretical analysis. Consequently, we have no convergence guarantees 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.

2.4 Related work

Performativity is a broad concept in the social sciences, philosophy, and economics [78, 115]. Below we focus on the relationship of our work to the most relevant technical scholarship.

A closely related line of work considers the problem of *concept drift*, broadly defined as the problem of learning when the target distribution over instances drifts with time. This setting has attracted attention both in the learning theory community [5, 6, 103] and by machine learning practitioners [64]. Concept drift is more general phenomenon than performativity in that it considers arbitrary sources of shift. However, studying the problem at this level of generality has led to a number of difficulties in creating a unified language and objective [64, 178], an issue we circumvent by assuming that the population distribution is determined by the deployed predictive model. Importantly, this line of work also discusses the importance of retraining [64, 174]. However, it stops short of discussing the need for stability or analyzing the long-term behavior of retraining.

Given that strategic classification is formally a special case of performative prediction, the study of performative optimality has been implicitly considered in the growing body of work on strategic classification [8, 76, 84, 126, 154]. More specifically, performatively optimal classifiers correspond to Stackelberg equilibria in strategic classification. In contrast to papers within this literature, our analysis relies on identifying macro-level assumptions on the loss and the distribution shift which make the problem tractable, rather than specific micro-level assumptions on the costs or utilities of the agents. For example, Dong et al. [49] prove that the institution's objective (performative risk) is convex by assuming that the agents are rational and compute best-responses according to particular utilities and cost functions. On the other hand, our conditions are on the distribution map and do not directly constrain behavior at the agent level.

The reader familiar with causality can think of $\mathcal{D}(\theta)$ as the interventional distribution over instances z resulting from a do-intervention that sets the model parameters to θ in some underlying causal graph. Importantly, this mapping $\mathcal{D}(\cdot)$ remains fixed and does not change over time or by intervention: deploying the same model at two different points in time must induce the same distribution over observations Z. While causal inference focuses on estimating properties of interventional distributions such as treatment effects [85, 135], our focus is on performative stability, performative optimality, and iterative retraining procedures. See the subsequent work of Mendler-Dünner et al. [122] for further connections to causal inference.

Finally, we would like to point out several works that have appeared subsequently to the work published in this thesis, but whose results contributed to the understanding of performative prediction. These include works studying stochastic optimization [43, 51, 182, 183], time-varying distribution shifts [20, 87, 110, 144], multi-player performative prediction [45, 111, 129, 138], and methods for finding optima [86, 116, 189], among others [32, 50, 74, 91, 96, 104, 117, 127, 139].

2.5 Deferred proofs

2.5.1 Auxiliary lemmas

Lemma 2.5.1 (Kantorovich-Rubinstein)**.** *A distribution map $\mathcal{D}(\cdot)$ is ε-sensitive if and only if for all $\theta, \theta' \in \Theta$:*

$$\sup\left\{\Big| \mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta)} g(z) - \mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta')} g(z) \Big| \leqslant \varepsilon\|\theta-\theta'\|_2 \ : \ g:\mathbb{R}^p\to\mathbb{R},\ g \textit{ 1-Lipschitz}\right\}.$$

Lemma 2.5.2. *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 $W_1(X,X')\leqslant C$. Then*

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

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

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

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 2.5.1 to obtain

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

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

Lemma 2.5.3. *Let $s \in (0,1)$, and fix $\alpha > 0$, then,*

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

Proof. Denote by $a_k \overset{\text{def}}{=} k^{-\alpha}$. Let $M_t = \max\{m \in \mathbb{N} : a_m > 2a_t\}$. 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 $a_k \leqslant 1$. For the second term, we use the fact that $a_k \leqslant 2a_t$ for $k > M_t$. We thus get:

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

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

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

□

2.5.2 Proof of Proposition 2.2.1

Proof of (a): Consider the linear loss defined as $\ell\left((x,y);\theta\right) = \beta y \theta$, for $\theta \in [-1,1]$. Note that this objective is β-jointly smooth and convex, but not strongly convex. Let the distribution of y according to $\mathcal{D}(\theta)$ be a point mass at $\varepsilon\theta$, 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 $\mathrm{DPR}(\theta, \varphi) = \varepsilon\beta\theta\varphi$. 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 $\varepsilon, \beta > 0$.

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 $\Theta \supseteq [-\frac{1}{2\varepsilon}, \frac{1}{2\varepsilon}]$.

Let the distribution of y according to $\mathcal{D}(\theta)$ be a point mass at $\varepsilon\theta$, 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 $\theta_1 = -\frac{1}{2\varepsilon}$. In the next step, again due to large C, we get $\theta_2 = 2$. Thus, RRM keeps oscillating between 2 and $-\frac{1}{2\varepsilon}$, failing to converge. This argument holds for all $\gamma, \varepsilon > 0$.

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 $1+\varepsilon\theta$, and let the distribution of x be invariant with respect to θ. This distribution family satisfies ε-sensitivity, because

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

By properties of the squared loss, we know

$$\arg\min_{\theta'} \mathrm{DPR}(\theta, \theta') = \mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta)}[y] = 1 + \varepsilon\theta.$$

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

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

which exactly matches the bound of Theorem 2.2.1 and proves the first statement of the proposition. The unique performatively stable point of this problem is θ such that $\theta = 1+\varepsilon\theta$, which is $\theta_{\mathrm{PS}} = \frac{1}{1-\varepsilon}$ for $\varepsilon > 1$.

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

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

thus concluding the proof.

2.5.3 Proof of Theorem 2.2.2

This proof is essentially a consequence of Lemma 2.2.1, proved in the following section. By following the steps of Lemma 2.2.1, we get

$$\begin{aligned}\|\theta_{t+1} - \theta_{\mathrm{PS}}\|_2^2 &\leqslant \|\theta_t - \theta_{\mathrm{PS}}\|_2^2 - 2\eta_t(\mathbb{E}\,\nabla\ell(z^{(t)};\theta_t))^\top(\theta_t - \theta_{\mathrm{PS}}) + \eta^2\|\,\mathbb{E}\,\nabla\ell(z^{(t)};\theta_t)\|_2^2 \\ &\overset{\text{def}}{=} B_1 - 2\eta B_2 + \eta^2 B_3,\end{aligned}$$

where we use $z^{(t)}$ to denote a sample from $\mathcal{D}(\theta_t)$.

Following the same approach as in Lemma 2.2.1, we get

$$B_2 \geqslant (\gamma - \varepsilon\beta)\|\theta_t - \theta_{\mathrm{PS}}\|_2^2.$$

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:

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

where in the third inequality we apply the fact that the loss if β-jointly smooth, together with Lemma 2.5.2. Putting everything together, this implies

$$\|\theta_{t+1} - \theta_{\mathrm{PS}}\|_2^2 \leqslant (1 - 2\eta(\gamma - \varepsilon\beta) + 2\eta^2\beta^2(1+\varepsilon^2))\|\theta_t - \theta_{\mathrm{PS}}\|_2^2.$$

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

$$\|\theta_{t+1} - \theta_{\mathrm{PS}}\|_2 \leqslant (1 - \eta(\gamma - \varepsilon\beta) + \eta^2\beta^2(1+\varepsilon^2))\|\theta_t - \theta_{\mathrm{PS}}\|_2.$$

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

$$\|\theta_{t+1} - \theta_{\mathrm{PS}}\|_2 \leqslant \left(1 - \frac{(\gamma-\varepsilon\beta)^2}{4(1+\varepsilon^2)\beta^2}\right)\|\theta_t - \theta_{\mathrm{PS}}\|_2.$$

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

We can unroll the above recursion to get

$$\begin{aligned}
\|\theta_{t+1} - \theta_{\mathrm{PS}}\|_2 &\leqslant \left(1 - \frac{(\gamma-\varepsilon\beta)^2}{4(1+\varepsilon^2)\beta^2}\right)^t\|\theta_1 - \theta_{\mathrm{PS}}\|_2 \\
&\leqslant \exp\left(-\frac{t(\gamma-\varepsilon\beta)^2}{4(1+\varepsilon^2)\beta^2}\right)\|\theta_1 - \theta_{\mathrm{PS}}\|_2.
\end{aligned}$$

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

2.5.4 Proof of Proposition 2.2.2

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

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

where $\beta \geqslant \gamma$ 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 $\theta_{\mathrm{PS}} = \frac{\beta/\gamma}{1-\varepsilon\beta/\gamma}$

whenever $\varepsilon \neq \frac{\gamma}{\beta}$; when $\varepsilon = \frac{\gamma}{\beta}$, there is no stable point. Repeated gradient descent has the dynamics:

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

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

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

2.5.5 Proof of Lemma 2.2.1

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

$$\|\theta_{t+1} - \theta_{\mathrm{PS}}\|_2^2 = \|\Pi_\Theta(\theta_t - \eta_t\nabla\ell(z^{(t)};\theta_t)) - \theta_{\mathrm{PS}}\|_2^2 \leqslant \|\theta_t - \eta_t\nabla\ell(z^{(t)};\theta_t) - \theta_{\mathrm{PS}}\|_2^2.$$

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

$$\begin{aligned}&\mathbb{E}\left[\|\theta_t - \eta_t\nabla\ell(z^{(t)};\theta_t) - \theta_{\mathrm{PS}}\|_2^2\right]\\ =\ &\mathbb{E}\left[\|\theta_t - \theta_{\mathrm{PS}}\|_2^2\right] - 2\eta_t\,\mathbb{E}\left[\nabla\ell(z^{(t)};\theta_t)^\top(\theta_t - \theta_{\mathrm{PS}})\right] + \eta_t^2\,\mathbb{E}\left[\|\nabla\ell(z^{(t)};\theta_t)\|_2^2\right]\\ \overset{\text{def}}{=}\ &B_1 - 2\eta_t B_2 + \eta_t^2 B_3.\end{aligned}$$

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

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

For the first term, we have that

$$\begin{aligned}&\mathbb{E}\left[(\nabla\ell(z^{(t)};\theta_t) - \nabla\ell(z^{(\theta_{\mathrm{PS}})};\theta_t)^\top(\theta_t - \theta_{\mathrm{PS}})\right]\\ &\quad= \mathbb{E}\left[\,\mathbb{E}\left[(\nabla\ell(z^{(t)};\theta_t) - \nabla\ell(z^{(\theta_{\mathrm{PS}})};\theta_t)^\top(\theta_t - \theta_{\mathrm{PS}}) \mid \theta_t\right]\,\right]\\ &\quad\geqslant -\varepsilon\beta\,\mathbb{E}\left[\|\theta_t - \theta_{\mathrm{PS}}\|_2^2\right].\end{aligned}$$

Having applied the law of iterated expectation, the above inequality follows from the fact that, conditional on θ_t, the function $\nabla\ell(z;\theta_t)^\top(\theta_t - \theta_{\mathrm{PS}})$ is $\beta\|\theta_t - \theta_{\mathrm{PS}}\|_2$—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 2.5.1 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,

$$\begin{aligned}&\mathbb{E}\left[(\nabla\ell(z^{(\theta_{\mathrm{PS}})};\theta_t) - \nabla\ell(z^{(\theta_{\mathrm{PS}})};\theta_{\mathrm{PS}}))^\top(\theta_t - \theta_{\mathrm{PS}})\right]\\ &\quad= \mathbb{E}\left[\,\mathbb{E}\left[(\nabla\ell(z^{(\theta_{\mathrm{PS}})};\theta_t) - \nabla\ell(z^{(\theta_{\mathrm{PS}})};\theta_{\mathrm{PS}}))^\top(\theta_t - \theta_{\mathrm{PS}}) \mid \theta_t\right]\,\right]\\ &\quad\geqslant \gamma\,\mathbb{E}\left[\|\theta_t - \theta_{\mathrm{PS}}\|_2^2\right].\end{aligned}$$

Therefore, we get that

$$B_2 \geqslant (\gamma - \varepsilon\beta)\,\mathbb{E}\left[\|\theta_t - \theta_{\mathrm{PS}}\|_2^2\right].$$

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

$$\begin{aligned}\mathbb{E}\left[\|\nabla\ell(z^{(t)};\theta_t)\|_2^2\right] &\leqslant \sigma^2 + L^2\,\mathbb{E}\left[\|\theta_t - G(\theta_t)\|_2^2\right]\\ &= \sigma^2 + L^2\,\mathbb{E}\left[\|\theta_t - \theta_{\mathrm{PS}} + \theta_{\mathrm{PS}} - G(\theta_t)\|_2^2\right]\\ &\leqslant \sigma^2 + L^2\left(\mathbb{E}\left[(\|\theta_t - \theta_{\mathrm{PS}}\|_2 + \|\theta_{\mathrm{PS}} - G(\theta_t)\|_2)^2\right]\right)\\ &\leqslant \sigma^2 + L^2\left(1 + \varepsilon\frac{\beta}{\gamma}\right)^2\mathbb{E}\left[\|\theta_t - \theta_{\mathrm{PS}}\|_2^2\right],\end{aligned}$$

where in the last step we use Theorem 2.2.1, which implies $\|\theta_{\mathrm{PS}} - G(\theta_t)\|_2 \leqslant \varepsilon\frac{\beta}{\gamma}\|\theta_t - \theta_{\mathrm{PS}}\|_2$.

Putting all the steps together completes the proof.

2.5.6 Proof of Theorem 2.2.3

From Lemma 2.2.1, we have that the following recursion holds:

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

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

$$\mathbb{E}\left[\|\theta_{t+1} - \theta_{\mathrm{PS}}\|_2^2\right] \leqslant \left(1 - 2\eta_t(\gamma - \varepsilon\beta) + 4\eta_t^2L^2\right)\mathbb{E}\left[\|\theta_t - \theta_{\mathrm{PS}}\|_2^2\right] + \eta_t^2\sigma^2.$$

We proceed by using induction. As in the theorem statement, we let $\eta_t = \frac{1}{(\gamma-\varepsilon\beta)(t+t_0)}$, where we denote $t_0 = \frac{8L^2}{(\gamma-\varepsilon\beta)^2}$. The base case, $t = 0$, is trivially true by construction of the bound and choice of t_0. Now, we adopt the inductive hypothesis that

$$\mathbb{E}\left[\|\theta_{t+1} - \theta_{\mathrm{PS}}\|_2^2\right] \leqslant \frac{\max\left\{2\sigma^2, 8L^2\|\theta_1 - \theta_{\mathrm{PS}}\|_2^2\right\}}{(\gamma - \varepsilon\beta)^2(t + t_0)}.$$

Then, by Lemma 2.2.1, it is true that

$$\begin{aligned}\mathbb{E}\left[\|\theta_{t+2}-\theta_{\mathrm{PS}}\|_2^2\right] &\leqslant \left(1-2\eta_t(\gamma-\varepsilon\beta)+4\eta_t^2L^2\right)\mathbb{E}\left[\|\theta_{t+1}-\theta_{\mathrm{PS}}\|_2^2\right]+\eta_t^2\sigma^2\\ &\leqslant \frac{1}{(\gamma-\varepsilon\beta)^2}\left(\frac{t+t_0-2+\frac{4L^2}{(\gamma-\varepsilon\beta)^2t_0}}{(t+t_0)^2}\max\left\{2\sigma^2,8L^2\|\theta_1-\theta_{\mathrm{PS}}\|_2^2\right\}+\frac{\sigma^2}{(t+t_0)^2}\right)\\ &\leqslant \frac{1}{(\gamma-\varepsilon\beta)^2}\left(\frac{t+t_0-1.5}{(t+t_0)^2}\max\left\{2\sigma^2,8L^2\|\theta_1-\theta_{\mathrm{PS}}\|_2^2\right\}+\frac{\sigma^2}{(t+t_0)^2}\right)\\ &\leqslant \frac{1}{(\gamma-\varepsilon\beta)^2}\left(\frac{t+t_0-1}{(t+t_0)^2}\max\left\{2\sigma^2,8L^2\|\theta_1-\theta_{\mathrm{PS}}\|_2^2\right\}-\frac{0.5\cdot 2\sigma^2-\sigma^2}{(t+t_0)^2}\right)\\ &= \frac{1}{(\gamma-\varepsilon\beta)^2}\cdot\frac{t+t_0-1}{(t+t_0)^2}\max\left\{2\sigma^2,8L^2\|\theta_1-\theta_{\mathrm{PS}}\|_2^2\right\}\\ &\leqslant \frac{1}{(\gamma-\varepsilon\beta)^2}\cdot\frac{1}{t+1+t_0}\max\left\{2\sigma^2,8L^2\|\theta_1-\theta_{\mathrm{PS}}\|_2^2\right\},\end{aligned}$$

where the last step follows because $(t+t_0)^2 > (t+t_0)^2-1 = (t+t_0+1)(t+t_0-1)$. Therefore, we have shown $\mathbb{E}\left[\|\theta_{t+2}-\theta_{\mathrm{PS}}\|_2^2\right] \leqslant \frac{M_{\mathrm{greedy}}}{(\gamma-\varepsilon\beta)^2(t+1+t_0)}$, which completes the proof by induction.

2.5.7 Proof of Theorem 2.2.4

To prove Theorem 2.2.4, we use the following classical result about convergence of SGD on a static distribution (see, e.g., [141]). The step size is chosen such that it matches the step size of Theorem 2.2.3 when $\varepsilon = 0$. We include the proof for completeness.

Lemma 2.5.4. *Under assumptions* (A1), (A2), *and* (A3), *lazy deploy satisfies the following:*

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

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

$$\mathbb{E}\left[\|\varphi_{t,j+1}-G(\theta_t)\|_2^2\right] \leqslant \frac{M_{\mathrm{lazy}}}{\gamma^2 j+L^2},$$

where $M_{\mathrm{lazy}} \overset{\mathrm{def}}{=} \max\{1.2\sigma^2, 8L^2\,\mathbb{E}[\|\theta_t-G(\theta_t)\|_2^2]\}$.

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

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

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

$$\begin{aligned}\mathbb{E}\left[\nabla\ell(z_j^{(t)};\varphi_{t,j})^\top(\varphi_{t,j} - G(\theta_t))\right] &\geqslant \mathbb{E}\left[(\nabla\ell(z_j^{(t)};\varphi_{t,j}) - \nabla\ell(z_j^{(t)}; G(\theta_t))^\top(\varphi_{t,j} - G(\theta_t))\right] \\ &\geqslant \gamma\,\mathbb{E}\left[\|\varphi_{t,j} - G(\theta_t)\|_2^2\right].\end{aligned}$$

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

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

Putting everything together, we get the desired recursion,

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

Now we turn to proving the second part of the lemma. Similarly to Theorem 2.2.3, we prove the result using induction. As in the theorem statement, we let $\eta_{t,j} = \frac{1}{\gamma(j+t_0)}$, where we denote $t_0 = \frac{8L^2}{\gamma^2}$. The base case, $j = 0$, is trivially true by construction of the bound and choice of t_0. Now, we adopt the inductive hypothesis that

$$\mathbb{E}\left[\|\varphi_{t,j+1} - G(\theta_t)\|_2^2\right] \leqslant \frac{\max\{1.2\sigma^2, 8L^2\,\mathbb{E}\left[\|\theta_t - G(\theta_t)\|_2^2\right]\}}{\gamma^2(j+t_0)}.$$

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

$$\begin{aligned}\mathbb{E}\left[\|\varphi_{t,j+2} - G(\theta_t)\|_2^2\right] &\leqslant \left(1 - 2\eta_{t,j}\gamma + \eta_{t,j}^2 L^2\right)\mathbb{E}\left[\|\varphi_{t,j+1} - G(\theta_t)\|_2^2\right] + \eta_{t,j}^2\sigma^2 \\ &\leqslant \frac{1}{\gamma^2}\left(\frac{j + t_0 - 2 + \frac{L^2}{\gamma^2 t_0}}{(j+t_0)^2}\max\left\{1.2\sigma^2, 8L^2\,\mathbb{E}\left[\|\theta_t - G(\theta_t)\|_2^2\right]\right\} + \frac{\sigma^2}{(j+t_0)^2}\right) \\ &\leqslant \frac{1}{\gamma^2}\left(\frac{j + t_0 - 15/8}{(j+t_0)^2}\max\left\{1.2\sigma^2, 8L^2\,\mathbb{E}\left[\|\theta_t - G(\theta_t)\|_2^2\right]\right\} + \frac{\sigma^2}{(j+t_0)^2}\right) \\ &\leqslant \frac{1}{\gamma^2}\left(\frac{j + t_0 - 1}{(j+t_0)^2}\max\left\{1.2\sigma^2, 8L^2\,\mathbb{E}\left[\|\theta_t - G(\theta_t)\|_2^2\right]\right\} - \frac{7/8\cdot 1.2\sigma^2 + \sigma^2}{(j+t_0)^2}\right) \\ &= \frac{1}{\gamma^2}\cdot\frac{j + t_0 - 1}{(j+t_0)^2}\max\left\{1.2\sigma^2, 8L^2\,\mathbb{E}\left[\|\theta_t - G(\theta_t)\|_2^2\right]\right\} \\ &\leqslant \frac{1}{\gamma^2}\cdot\frac{1}{j+1+t_0}\max\left\{1.2\sigma^2, 8L^2\,\mathbb{E}\left[\|\theta_t - G(\theta_t)\|_2^2\right]\right\},\end{aligned}$$

where the last step follows because $(j + t_0)^2 > (j + t_0)^2 - 1 = (j + t_0 + 1)(j + t_0 - 1)$. Therefore, we have shown $\mathbb{E}\left[\|\varphi_{t,j+2} - G(\theta_t)\|_2^2\right] \leqslant \frac{M_{\text{lazy}}}{\gamma^2(j+1+t_0)}$, which completes the proof by induction. □

Now we prove Theorem 2.2.4.

First we state two identities used in the proof, which follow from Theorem 2.2.1:

$$\|G(\theta) - \theta_{\text{PS}}\|_2 \leqslant \varepsilon\frac{\beta}{\gamma}\|\theta - \theta_{\text{PS}}\|_2, \tag{2.9}$$

$$\|\theta - G(\theta)\|_2 \leqslant \|\theta - \theta_{\text{PS}}\|_2 + \|\theta_{\text{PS}} - G(\theta)\|_2 \leqslant \left(1 + \varepsilon\frac{\gamma}{\beta}\right)\|\theta - \theta_{\text{PS}}\|_2. \tag{2.10}$$

Note that identity (2.10) implies $\|\theta - G(\theta)\|_2 < 2\|\theta - \theta_{\text{PS}}\|_2$ if $\varepsilon < \frac{\gamma}{\beta}$.

By triangle inequality, we have

$$\begin{aligned}
&\mathbb{E}\left[\|\theta_{t+1} - \theta_{\text{PS}}\|_2^2\right] \\
&\quad = \mathbb{E}\left[\|\theta_{t+1} - G(\theta_t) + G(\theta_t) - \theta_{\text{PS}}\|_2^2\right] \\
&\quad \leqslant \mathbb{E}\left[\|\theta_{t+1} - G(\theta_t)\|_2^2\right] + 2\,\mathbb{E}\left[\|\theta_{t+1} - G(\theta_t)\|_2\|G(\theta_t) - \theta_{\text{PS}}\|_2\right] + \mathbb{E}\left[\|G(\theta_t) - \theta_{\text{PS}}\|_2^2\right].
\end{aligned} \tag{2.11}$$

Denoting $t_0 = \frac{8L^2}{\gamma^2}$, Lemma 2.5.4 bounds the first term by

$$\begin{aligned}
\mathbb{E}\left[\|\theta_{t+1} - G(\theta_t)\|_2^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\|\theta_{t+1} - G(\theta_t)\|_2^2 \mid \theta_t\right]\right] \\
&\leqslant \frac{1.2\sigma^2 + 8L^2\,\mathbb{E}\left[\|\theta_t - G(\theta_t)\|_2^2\right]}{\gamma^2(n(t) + t_0)} \\
&\leqslant \frac{1.2\sigma^2 + 32L^2\,\mathbb{E}\left[\|\theta_t - \theta_{\text{PS}}\|_2^2\right]}{\gamma^2(n(t) + t_0)},
\end{aligned}$$

where in the last step we apply identity (2.10). Note also that by Jensen's inequality, we know

$$\mathbb{E}\left[\|\theta_{t+1} - G(\theta_t)\|_2\right] \leqslant \frac{1.1\sigma + 6L\,\mathbb{E}\left[\|\theta_t - G(\theta_t)\|_2\right]}{\gamma\sqrt{n(t) + t_0}}.$$

We can use this inequality, together with identities (2.9) and (2.10), to bound the cross-term in equation (2.11) as follows:

$$\begin{aligned}
&2\,\mathbb{E}\left[\|\theta_{t+1} - G(\theta_t)\|_2\|G(\theta_t) - \theta_{\text{PS}}\|_2\right] \\
\leqslant\;& 2\varepsilon\frac{\beta}{\gamma}\,\mathbb{E}\left[\|\theta_{t+1} - G(\theta_t)\|_2\|\theta_t - \theta_{\text{PS}}\|_2\right] \\
\leqslant\;& \frac{2\varepsilon\frac{\beta}{\gamma}}{\sqrt{n(t) + t_0}}\,\mathbb{E}\left[\left(\frac{6L}{\gamma}\|\theta_t - G(\theta_t)\|_2 + \frac{1.1\sigma}{\gamma}\right)\|\theta_t - \theta_{\text{PS}}\|_2\right] \\
\leqslant\;& \frac{2\varepsilon\frac{\beta}{\gamma}}{\sqrt{n(t) + t_0}}\,\mathbb{E}\left[\left(\frac{6L}{\gamma}\left(1 + \varepsilon\frac{\beta}{\gamma}\right)\|\theta_t - \theta_{\text{PS}}\|_2 + \frac{1.1\sigma}{\gamma}\right)\|\theta_t - \theta_{\text{PS}}\|_2\right] \\
\leqslant\;& \frac{24\varepsilon\beta L}{\gamma^2\sqrt{n(t) + t_0}}\,\mathbb{E}\left[\|\theta_t - \theta_{\text{PS}}\|_2^2\right] + \frac{2.2\sigma\varepsilon\beta}{\gamma^2\sqrt{n(t) + t_0}}\,\mathbb{E}\left[\|\theta_t - \theta_{\text{PS}}\|_2\right].
\end{aligned}$$

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

$$\frac{2.2\sigma\varepsilon\beta}{\gamma^2\sqrt{n(t)+t_0}}\,\mathbb{E}\left[\|\theta_t - \theta_{\mathrm{PS}}\|_2\right] \leqslant \frac{1.1\sigma\varepsilon\beta}{\gamma^2}\left(\frac{1}{(n(t)+t_0)^{\alpha_0}} + \frac{\mathbb{E}\left[\|\theta_t - \theta_{\mathrm{PS}}\|_2^2\right]}{(n(t)+t_0)^{1-\alpha_0}}\right).$$

Thus, the final bound on the cross-term in equation (2.11) is

$$\begin{aligned} 2\,\mathbb{E}\left[\|\theta_{t+1} - G(\theta_t)\|_2\|G(\theta_t) - \theta_{\mathrm{PS}}\|_2\right] \leqslant &\left(\frac{24\varepsilon\beta L}{\gamma^2\sqrt{n(t)+t_0}} + \frac{1.1\sigma\varepsilon\beta}{\gamma^2(n(t)+t_0)^{1-\alpha_0}}\right)\mathbb{E}\left[\|\theta_t - \theta_{\mathrm{PS}}\|_2^2\right] \\ &+ \frac{1.1\sigma\varepsilon\beta}{\gamma^2(n(t)+t_0)^{\alpha_0}}. \end{aligned}$$

The final term in equation (2.11) can be bounded by identity (2.9):

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

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

$$\begin{aligned} \mathbb{E}\left[\|\theta_{t+1} - \theta_{\mathrm{PS}}\|_2^2\right] \leqslant &\left(\frac{32L^2}{\gamma^2(n(t)+t_0)} + \frac{24\varepsilon\beta L}{\gamma^2\sqrt{n(t)+t_0}} + \frac{1.1\sigma\varepsilon\beta}{\gamma^2(n(t)+t_0)^{1-\alpha_0}} + \left(\varepsilon\frac{\beta}{\gamma}\right)^2\right)\mathbb{E}\left[\|\theta_t - \theta_{\mathrm{PS}}\|_2^2\right] \\ &+ \frac{1.2\sigma^2}{\gamma^2(n(t)+t_0)} + \frac{1.1\sigma\varepsilon\beta}{\gamma^2(n(t)+t_0)^{\alpha_0}} \\ \leqslant\ & c\,\mathbb{E}\left[\|\theta_t - \theta_{\mathrm{PS}}\|_2^2\right] + \frac{1.2\sigma^2}{\gamma^2(n(t)+t_0)} + \frac{1.1\sigma\varepsilon\beta}{\gamma^2(n(t)+t_0)^{\alpha_0}}, \end{aligned} \tag{2.12}$$

where we define

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

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

Unrolling the recursion given by equation (2.12), we get

$$\mathbb{E}\left[\|\theta_{t+1} - \theta_{\mathrm{PS}}\|_2^2\right] \leqslant c^t\|\theta_1 - \theta_{\mathrm{PS}}\|_2^2 + \frac{1}{\gamma^2}\sum_{j=1}^{t} c^{t-j}\left(\frac{1.2\sigma^2}{n(j)+t_0} + \frac{1.1\sigma\varepsilon\beta}{(n(j)+t_0)^{\alpha_0}}\right).$$

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

$$\begin{aligned} &\frac{1}{\gamma^2}\sum_{j=1}^{t} c^{t-j}\left(\frac{1.2\sigma^2}{n(j)+t_0} + \frac{1.1\sigma\varepsilon\beta}{(n(j)+t_0)^{\alpha_0}}\right) \\ \leqslant\ & \frac{1.2\sigma^2}{\gamma^2}\sum_{j=1}^{t} c^{t-j}\frac{1}{n(j)+t_0} + \frac{1.1\sigma\varepsilon\beta}{\gamma^2}\sum_{j=1}^{t} c^{t-j}\frac{1}{(n(j)+t_0)^{\alpha_0}} \\ \leqslant\ & \frac{1}{\gamma^2(1-c)}\left(\frac{1.2\sigma^2}{n_0}(2t^{-\alpha} + c^{(1-2^{-1/\alpha})t}) + \frac{1.1\sigma\varepsilon\beta}{n_0^{\alpha_0}}(2t^{-\alpha\cdot\alpha_0} + c^{(1-2^{-1/(\alpha\alpha_0)})t})\right) \end{aligned}$$

where in the second inequality we apply Lemma 2.5.3 after plugging in the choice of $n(t)$. Using the fact that $\alpha_0 \in (0,1)$ and hence $c^{(1-2^{-1/(\alpha\alpha_0)})t} < c^{(1-2^{-1/\alpha})t}$, as well as $\varepsilon < \frac{\gamma}{\beta}$ and $n_0 \geqslant 1$, gives

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

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

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

as desired.

2.5.8 Proof of Theorem 2.3.1

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

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

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

$$\left(\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta)}[\nabla_\theta\ell(z;\theta) + \ell(z;\theta)\nabla_\theta\log(p_\theta(z))]\right)^\top(\theta' - \theta) + \frac{\gamma - 2\varepsilon\beta}{2}\|\theta - \theta'\|_2^2 \leqslant \mathrm{PR}(\theta') - \mathrm{PR}(\theta), \tag{2.13}$$

for all $\theta, \theta' \in \Theta$. By assumption (A5), we know that for all $\theta, \theta', \theta_0 \in \Theta$,

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

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:

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

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

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

Going back to equation (2.13), we see that a sufficient condition for $(\gamma-2\varepsilon\beta)$-convexity of the performative risk is

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

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

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

and thus we have further simplified the sufficient condition to

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

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 (Lemma 2.5.1):

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta)}[\nabla_\theta\ell(z;\theta)]^\top(\theta'-\theta) - \mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta')}[\nabla_\theta\ell(z;\theta)]^\top(\theta'-\theta) \leqslant \varepsilon\beta\|\theta-\theta'\|_2^2. \tag{2.14}$$

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

2.5.9 Proof of Theorem 2.3.2

Following the steps of Theorem 2.3.1, we know that $\mathrm{PR}(\theta)$ is λ-convex if and only if

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

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

We now state a technical lemma, deferring its proof to the end of this section.

Lemma 2.5.5. *Suppose that*

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\alpha\theta+(1-\alpha)\theta')}[g(z)] \leqslant \mathop{\mathbb{E}}_{z\sim\alpha\mathcal{D}(\theta)+(1-\alpha)\mathcal{D}(\theta')}[g(z)] - \frac{\alpha(1-\alpha)\gamma_z}{2}\mathbb{E}\|\Sigma(\theta-\theta')z_0 + \mu(\theta-\theta')\|_2^2.$$

Then,

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta')}[g(z)] \geqslant \mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta)}[g(z)] + (\nabla_\theta \mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta)}[g(z)])^\top(\theta'-\theta) + \frac{\gamma_z}{2}\mathbb{E}\|\Sigma(\theta-\theta')z_0 + \mu(\theta-\theta')\|_2^2.$$

Note that the assumption of the lemma is implied by the location-family form of the distribution map, together with strong convexity of the loss in z. Therefore, by Lemma 2.5.5, we know

$$\begin{aligned}\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta)}[\ell(z;\theta)\nabla_\theta\log(p_\theta(z))]^\top(\theta'-\theta) \leqslant \mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta')}[\ell(z;\theta)] - \mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta)}[\ell(z;\theta)]\\ - \frac{\gamma_z}{2}\mathbb{E}\,\|\Sigma(\theta-\theta')z_0+\mu(\theta-\theta')\|_2^2,\end{aligned}$$

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

Thus it suffices to show

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta)}[\nabla_\theta\ell(z;\theta)]^\top(\theta'-\theta)+\frac{\lambda}{2}\|\theta-\theta'\|_2^2 \leqslant \mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta')}\ell(z;\theta')-\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta')}\ell(z;\theta)+\frac{\gamma_z}{2}\mathbb{E}\,\|\Sigma(\theta-\theta')z_0+\mu(\theta-\theta')\|_2^2.$$

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

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

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

$$(\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta)}[\nabla_\theta\ell(z;\theta)] - \mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta')}[\nabla_\theta\ell(z;\theta)])^\top(\theta'-\theta) \leqslant \frac{\gamma-\lambda}{2}\|\theta-\theta'\|_2^2+\frac{\gamma_z}{2}\mathbb{E}\,\|\Sigma(\theta-\theta')z_0+\mu(\theta-\theta')\|_2^2. \tag{2.15}$$

We bound the left-hand side by applying smoothness of the loss together with the Kantorovich-Rubinstein duality (Lemma 2.5.1); for this, we need a bound on $W_1(\mathcal{D}(\theta),\mathcal{D}(\theta'))$. We will use the bound implied by ε-sensitivity, as well as the bound implied by the following lemma.

Lemma 2.5.6. *Suppose that the distribution map $\mathcal{D}(\theta)$ forms a location-scale family* (2.5)*. Then,*

$$W_1(\mathcal{D}(\theta),\mathcal{D}(\theta')) \leqslant \mathbb{E}\,\|\Sigma(\theta-\theta')z_0+\mu(\theta-\theta')\|_2.$$

Proof of Lemma 2.5.6. By definition,

$$W_1(\mathcal{D}(\theta),\mathcal{D}(\theta')) = \inf_{\Pi(\mathcal{D}(\theta),\mathcal{D}(\theta'))}\mathop{\mathbb{E}}_{(z_\theta,z_{\theta'})\sim\Pi(\mathcal{D}(\theta),\mathcal{D}(\theta'))}[\|z_\theta-z_{\theta'}\|_2],$$

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 $z_0\sim\mathcal{D}$, and set $z_\theta=(\Sigma_0+\Sigma(\theta))z_0+\mu_0+\mu(\theta)$ and $z_{\theta'}=(\Sigma_0+\Sigma(\theta'))z_0+\mu_0+\mu(\theta')$. With this choice, $\|z_\theta-z_{\theta'}\|_2=\|\Sigma(\theta-\theta')z_0+\mu(\theta-\theta')\|_2$, and hence $W_1(\mathcal{D}(\theta),\mathcal{D}(\theta'))\leqslant\mathbb{E}\,\|\Sigma(\theta-\theta')z_0+\mu(\theta-\theta')\|_2$. □

Therefore, the left-hand side in equation (2.15) can be bounded by

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta)}[\nabla_\theta\ell(z;\theta)]^\top(\theta'-\theta)-\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta')}[\nabla_\theta\ell(z;\theta)]^\top(\theta'-\theta) \leqslant \beta\,\mathbb{E}\,\|\Sigma(\theta-\theta')z_0+\mu(\theta-\theta')\|_2\|\theta'-\theta\|_2,$$

but also by applying ε-sensitivity

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

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

$$\beta\,\mathbb{E}\,\|\Sigma(\theta-\theta')z_0 + \mu(\theta-\theta')\|_2\|\theta'-\theta\|_2 \leqslant \frac{\beta^2/\gamma_z}{2}\|\theta-\theta'\|_2^2 + \frac{\gamma_z}{2}\,\mathbb{E}\,\|\Sigma(\theta-\theta')z_0 + \mu(\theta-\theta')\|_2^2 \tag{2.16}$$

and

$$\beta\varepsilon\|\theta'-\theta\|_2^2 \leqslant \frac{2\beta\varepsilon - \gamma_z(\sigma_{\min}^2(\mu)+\sigma_{\min}^2(\Sigma))}{2}\|\theta-\theta'\|_2^2 + \frac{\gamma_z}{2}\,\mathbb{E}\,\|\Sigma(\theta-\theta')z_0 + \mu(\theta-\theta')\|_2^2. \tag{2.17}$$

By the AM-GM inequality, we have

$$\beta\,\mathbb{E}\,\|\Sigma(\theta-\theta')z_0 + \mu(\theta-\theta')\|_2\|\theta'-\theta\|_2 \leqslant \frac{1}{2}\frac{\beta^2}{\gamma_z}\|\theta'-\theta\|_2^2 + \frac{\gamma_z}{2}\,\mathbb{E}\,\|\Sigma(\theta-\theta')z_0 + \mu(\theta-\theta')\|_2^2,$$

and so condition (2.16) follows.

For condition (2.17), we observe that

$$\begin{aligned}\mathbb{E}\,\|\Sigma(\theta-\theta')z_0 + \mu(\theta-\theta')\|_2^2 &= \mathbb{E}\,\|\Sigma(\theta-\theta')z_0\|_2^2 + \|\mu(\theta-\theta')\|_2^2\\ &= \mathrm{Tr}\left(\Sigma(\theta-\theta')\Sigma_{z_0}\Sigma(\theta-\theta')^\top\right) + \|\mu(\theta-\theta')\|_2^2\\ &= \|\Sigma_{z_0}^{1/2}\Sigma(\theta-\theta')^\top\|_F^2 + \|\mu(\theta-\theta')\|_2^2.\end{aligned}$$

Applying $\sigma_{\min}(\Sigma)\|\theta-\theta'\|_2 \leqslant \|\Sigma_{z_0}^{1/2}\Sigma(\theta-\theta')^\top\|_F$ and $\sigma_{\min}(\mu)\|\theta-\theta'\|_2 \leqslant \|\mu(\theta-\theta')\|_2$ completes the proof of the theorem.

Proof of Lemma 2.5.5. The proof follows the standard argument for proving equivalent formulations of strong convexity.

First we show that $\mathbb{E}_{z\sim\mathcal{D}(\theta)}[g(z)]-\frac{\gamma_z}{2}\,\mathbb{E}\,\|\Sigma(\theta)z_0+\mu\theta\|_2^2$ is convex in θ. This follows because:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\alpha\theta+(1-\alpha)\theta')}[g(z)]-\frac{\gamma_z}{2}\,\mathbb{E}\,\|\Sigma(\alpha\theta+(1-\alpha)\theta')z_0+\mu(\alpha\theta+(1-\alpha)\theta')\|_2^2\\
&\quad\leqslant \mathop{\mathbb{E}}_{z\sim\alpha\mathcal{D}(\theta)+(1-\alpha)\mathcal{D}(\theta')}[g(z)]-\frac{\alpha(1-\alpha)\gamma_z}{2}\,\mathbb{E}\,\|\Sigma(\theta-\theta')z_0+\mu(\theta-\theta')\|_2^2\\
&\quad-\frac{\gamma_z}{2}\,\mathbb{E}\,\|\Sigma(\alpha\theta+(1-\alpha)\theta')z_0+\mu(\alpha\theta+(1-\alpha)\theta')\|_2^2\\
&\quad= \mathop{\mathbb{E}}_{z\sim\alpha\mathcal{D}(\theta)+(1-\alpha)\mathcal{D}(\theta')}[g(z)]-\frac{\gamma_z}{2}\alpha^2\,\mathbb{E}\,\|\Sigma(\theta)z_0+\mu\theta\|_2^2-\frac{\gamma_z}{2}(1-\alpha)^2\,\mathbb{E}\,\|\Sigma(\theta')z_0+\mu\theta'\|_2^2\\
&\quad+\frac{\gamma_z}{2}2\alpha(1-\alpha)\,\mathbb{E}(\Sigma(\theta)+\mu\theta)^\top(\Sigma(\theta')+\mu\theta')-\frac{\alpha(1-\alpha)\gamma_z}{2}\,\mathbb{E}\,\|\Sigma(\theta-\theta')z_0+\mu(\theta-\theta')\|_2^2\\
&\quad= \mathop{\mathbb{E}}_{z\sim\alpha\mathcal{D}(\theta)+(1-\alpha)\mathcal{D}(\theta')}[g(z)]-\frac{\gamma_z}{2}\alpha\,\mathbb{E}\,\|\Sigma(\theta)z_0+\mu\theta\|_2^2-\frac{\gamma_z}{2}(1-\alpha)\,\mathbb{E}\,\|\Sigma(\theta')z_0+\mu\theta'\|_2^2\\
&\quad=\alpha\left(\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta)}[g(z)]-\frac{\gamma_z}{2}\,\mathbb{E}\,\|\Sigma(\theta)z_0+\mu\theta\|_2^2\right)-(1-\alpha)\left(\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta')}[g(z)]\frac{\gamma_z}{2}\,\mathbb{E}\,\|\Sigma(\theta')z_0+\mu\theta'\|_2^2\right).
\end{aligned}$$

By the equivalent first-order characterization, this means that

$$\begin{aligned}
\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta')}[g(z)]&\geqslant\frac{\gamma_z}{2}\,\mathbb{E}\,\|\Sigma(\theta')z_0+\mu\theta'\|_2^2+\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta)}[g(z)]-\frac{\gamma_z}{2}\,\mathbb{E}\,\|\Sigma(\theta)z_0+\mu\theta\|_2^2\\
&\quad+(\nabla_\theta\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta)}[g(z)])^\top(\theta'-\theta)-\frac{\gamma_z}{2}2\,\mathbb{E}(\Sigma(\theta)z_0+\mu\theta)^\top(\nabla_\theta(\Sigma(\theta)z_0+\mu\theta))^\top(\theta'-\theta)\\
&\geqslant\frac{\gamma_z}{2}\,\mathbb{E}\,\|\Sigma(\theta')z_0+\mu\theta'\|_2^2+\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta)}[g(z)]-\frac{\gamma_z}{2}\,\mathbb{E}\,\|\Sigma(\theta)z_0+\mu\theta\|_2^2\\
&\quad+(\nabla_\theta\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta)}[g(z)])^\top(\theta'-\theta)-\gamma_z\,\mathbb{E}(\Sigma(\theta)z_0+\mu\theta)^\top(\Sigma(\theta'-\theta)z_0+\mu(\theta'-\theta))\\
&=\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta)}[g(z)]+(\nabla_\theta\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta)}[g(z)])^\top(\theta'-\theta)+\frac{\gamma_z}{2}\,\mathbb{E}\,\|\Sigma(\theta-\theta')z_0+\mu(\theta-\theta')\|_2^2.
\end{aligned}$$

□

Remark 2.5.1. *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* (2.17)*, we saw that*

$$\mathbb{E}\,\|\Sigma(\theta-\theta')z_0+\mu(\theta-\theta')\|_2^2=\|\Sigma_{z_0}^{1/2}\Sigma(\theta-\theta')^\top\|_F^2+\|\mu(\theta-\theta')\|_2^2.$$

If we then denote

$$\sigma_{\max}(\mu)=\max_{\|\theta\|_2=1}\|\mu\theta\|_2,\quad \sigma_{\max}(\Sigma)=\max_{\|\theta\|_2=1}\|\Sigma_{z_0}^{1/2}\Sigma(\theta)^\top\|_F,$$

we can see that $\mathbb{E}\,\|\Sigma(\theta-\theta')z_0+\mu(\theta-\theta')\|_2^2\leqslant\sigma_{\max}^2(\mu)\|\theta-\theta'\|_2^2+\sigma_{\max}^2(\Sigma)\|\theta-\theta'\|_2^2$. *Combining this result with Lemma 2.5.6 and Jensen's inequality, we get that*

$$W_1(\mathcal{D}(\theta),\mathcal{D}(\theta'))\leqslant\sqrt{\sigma_{\max}^2(\mu)+\sigma_{\max}^2(\Sigma)}\|\theta-\theta'\|_2,$$

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

2.5.10 Proof of Theorem 2.3.3

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

$$z_\theta \sim \mathcal{D}(\theta) \;\Leftrightarrow\; z_\theta \overset{d}{=} z_0 + \mu\theta,$$

where $z_0 \sim \mathcal{D}_0$. 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 $\lambda = \max\{\gamma - \beta^2/\gamma_z, \gamma - 2\varepsilon\beta + \gamma_z\sigma^2_{\min}(\mu)\} > 0$, where γ and γ_z are the strong convexity parameters of the loss in θ and z, respectively. By Theorem 2.3.2, 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 $\Theta = \{\theta : \|\theta\|_2 \leqslant R\}$. Finally, we assume that the minimizer of the perturbed performative risk at the population level, $\hat\theta \in \arg\min_{\theta\in\Theta} \widehat{\mathrm{PR}}(\theta)$ is contained in the interior of Θ with probability 1.

Theorem 2.5.1. *Under the preceding assumptions, if* $n \geqslant \Omega\left(d + m + \log(1/\delta)\right)$*, then, with probability* $1-\delta$*, Algorithm 1 returns a point* $\hat\theta_n$ *such that*

$$\mathrm{PR}(\hat\theta_n) - \mathrm{PR}(\theta_{\mathrm{PO}}) \leqslant \mathcal{O}\left(\frac{d + m + \log(1/\delta)}{n} + \frac{1}{\delta n}\right).$$

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

Lemma 2.5.7. *If* $n \geqslant \Omega(d + m + \log(1/\delta))$*, then with probability* $1-\delta$*,*

$$\|\mu - \hat\mu\|_2 \leqslant O\left(\sqrt{\frac{(d+m) + \log(1/\delta)}{n}}\right).$$

The next lemma is a simple adaptation from Theorem 2 in [152] controlling the generalization gap of the empirical risk minimizer for strongly convex losses.

Lemma 2.5.8. *Suppose* $\widehat{\mathrm{PR}}_n$ *is* $\hat\lambda$*-strongly convex. Then, with probability at least* $1-\delta$*,*

$$\widehat{\mathrm{PR}}(\hat\theta_n) - \widehat{\mathrm{PR}}(\hat\theta) \leqslant \frac{4(L_z\|\hat\mu\|_2 + L)^2}{\delta\hat\lambda n}.$$

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

Lemma 2.5.9. *For any* $\theta \in \Theta$*,*

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

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 2.5.10. *Under the proceeding assumptions, the performative risk* $\mathrm{PR}(\theta)$ *is* $\beta_\theta = O(\|\mu\|_2^2)$*-smooth.*

With these lemmas in hand, we are now ready to prove Theorem 2.5.1.

Proof of Theorem 2.5.1. By assumption, the performative risk $\mathrm{PR}(\theta)$ is λ-strongly convex, for some $\lambda > 0$. This implies

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

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 $\|a+b\|_2^2 \leqslant 2\|a\|_2^2 + 2\|b\|_2^2$,

$$\begin{aligned}
\frac{1}{2\lambda}\|\nabla \mathrm{PR}(\hat{\theta}_n)\|_2^2 &= \frac{1}{2\lambda}\|\nabla \mathrm{PR}(\hat{\theta}_n) - \nabla\widehat{\mathrm{PR}}(\hat{\theta}_{\mathrm{PO}})\|_2^2 \\
&= \frac{1}{2\lambda}\|\nabla \mathrm{PR}(\hat{\theta}_n) - \nabla\widehat{\mathrm{PR}}(\hat{\theta}_n) + \nabla\widehat{\mathrm{PR}}(\hat{\theta}_n) - \nabla\widehat{\mathrm{PR}}(\hat{\theta}_{\mathrm{PO}})\|_2^2 \\
&\leqslant \frac{1}{\lambda}\|\nabla \mathrm{PR}(\hat{\theta}_n) - \nabla\widehat{\mathrm{PR}}(\hat{\theta}_n)\|_2^2 + \frac{1}{\lambda}\|\nabla\widehat{\mathrm{PR}}(\hat{\theta}_n) - \nabla\widehat{\mathrm{PR}}(\hat{\theta}_{\mathrm{PO}})\|_2^2. \qquad (2.18)
\end{aligned}$$

We bound each of these terms separately. For the first term, by Lemma 2.5.9,

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

By Lemma 2.5.7, with probability $1-\delta$, we can bound $\|\mu - \hat{\mu}\|_2^2 \leqslant O\left(\frac{d+m+\log(1/\delta)}{n}\right)$, and thus

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

For the second term in equation (2.18), notice that $\lambda = \max\{\gamma - \beta^2/\gamma_z, \gamma - 2\varepsilon\beta + \gamma_z\sigma_{\min}^2(\mu)\} > 0$ implies that $\widehat{\mathrm{PR}}$ is at least $\hat{\lambda} = \lambda - O(\frac{1}{\sqrt{n}})$-strongly convex. This follows because $|\sigma_{\min}(\mu) - \sigma_{\min}(\hat{\mu})| \leqslant \|\mu - \hat{\mu}\|$ by Weyl's inequality, and $\widehat{\mathrm{PR}}$ is $O(\|\hat{\mu}\|_2)$-sensitive, so by Lemma 2.5.7, 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 $n \geqslant \Omega(1/\lambda^2)$, 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 $n \geqslant \Omega(1/\lambda^2)$,

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

Now, when $\widehat{\mathrm{PR}}$ is strongly convex, the finite-sample performative risk $\widehat{\mathrm{PR}}_n$ is also strongly convex because Theorem 2.3.2 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 $\{z_1, \dots, z_n\}$. Consequently, by Lemma 2.5.8, with probability $1-\delta$,

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

By Lemma 2.5.10, $\widehat{\mathrm{PR}}$ is $O(\|\hat{\mu}\|_2^2)$-smooth. Applying the previous display then gives us,

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

By the triangle inequality and repeated application of $(a+b)^2 \leqslant 2a^2+2b^2$, $\|\hat{\mu}\|_2^6 \leqslant 128\|\hat{\mu}-\mu\|_2^6 + 128\|\mu\|_2^6$. Therefore, the above term is $O(\|\mu\|_2^6/\delta n)$. Putting everything together with a union bound, we have shown that with probability $1-\delta$, if $n \geqslant \Omega(d+m+\log(1/\delta))$, it holds that

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

as desired. □

Proofs of technical lemmas. The proof of Lemma 2.5.7 is essentially standard (see, e.g., [120]), but we include it for completeness.

Proof of Lemma 2.5.7. Define $Z \in \mathbb{R}^{n\times m}$ with rows z_i and $\Theta \in \mathbb{R}^{n\times d}$ with rows θ_i, $1 \leqslant i \leqslant n$. 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

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

Consequently, we can bound the estimation error as

$$\begin{aligned}\|\mu - \hat{\mu}\|_2 = \|\mu^\top - \hat{\mu}^\top\|_2 &= \|\mu^\top - \left(\Theta^\top\Theta\right)^{-1}\Theta^\top\left(\Theta\mu^\top + Z_0\right)\|_2 \\ &= \|\left(\Theta^\top\Theta\right)^{-1}\Theta^\top Z_0\|_2 \\ &\leqslant \frac{1}{\lambda_{\min}(\Theta^\top\Theta)}\|\Theta^\top Z_0\|_2.\end{aligned}$$

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 [168]) gives, with probability $1-\delta$,

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

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

$$\sqrt{\lambda_{\min}(\Theta^\top \Theta)} \geqslant \Omega(\sqrt{n}/2). \tag{2.19}$$

Control of the second term, $\|\Theta^\top Z_0\|_2$, also follows from a standard covering argument followed by the Bernstein bound. Write $\Theta^\top Z_0 = \sum_{i=1}^n \theta_i (z_0)_i^\top$. Let $\mathcal{B}^d$ and $\mathcal{B}^m$ denote the unit balls in $\mathbb{R}^d$ and $\mathbb{R}^m$, respectively. Then,

$$\|\Theta^\top Z_0\|_2 = \sup_{x \in \mathcal{B}^d, y \in \mathcal{B}^m} x^\top \left(\sum_{i=1}^n \theta_i (z_0)_i^\top \right) y = \sup_{x \in \mathcal{B}^d, y \in \mathcal{B}^m} \sum_{i=1}^n \left(x^\top \theta_i \right) \left((z_0)_i^\top y \right).$$

Let $\mathcal{N}_\varepsilon$, and $\mathcal{M}_\varepsilon$ denote ε-coverings of $\mathcal{B}^d$ and $\mathcal{B}^m$, respectively. A volumetric bound gives $|\mathcal{N}_\varepsilon| \leqslant \left(1 + \frac{2}{\varepsilon}\right)^d$ and similarly $|\mathcal{M}_\varepsilon| \leqslant \left(1 + \frac{2}{\varepsilon}\right)^m$ (see Corollary 4.2.13 in [168]). Taking $\varepsilon = 1/4$, $|\mathcal{N}_\varepsilon| \leqslant 9^d$ and $|\mathcal{M}_\varepsilon| \leqslant 9^m$. Approximating the supremum over the ε-nets gives

$$\|\Theta^\top Z_0\|_2 \leqslant 2 \max_{x \in \mathcal{N}_\varepsilon, y \in \mathcal{M}_\varepsilon} \sum_{i=1}^n \left(x^\top \theta_i \right) \left((z_0)_i^\top y \right).$$

Fix $x, y \in \mathcal{N}_\varepsilon, \mathcal{M}_\varepsilon$. 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 $(z_0)_i$ is subgaussian with parameter K and $\|y\|_2 = 1$, the marginal $(z_0)_i^\top y$ is subgaussian with parameter K. Since z_0 and θ are independent and zero-mean, the product $(x^\top \theta_i)((z_0)_i^\top y)$ is zero-mean and subexponential with parameter K. Since each term is subexponential, by the Bernstein bound (see Theorem 2.8.1 in [168]), for any $t > 0$,

$$\mathbb{P}\left\{ \sum_{i=1}^n \left(x^\top \theta_i \right) \left((z_0)_i^\top y \right) > t/2 \right\} \leqslant \exp\left(-c \min\left\{ \frac{t^2}{nK^2}, \frac{t}{K} \right\} \right),$$

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

$$\mathbb{P}\left\{ \|\Theta^\top Z_0\|_2 > t \right\} \leqslant 9^{d+m} \exp\left(-c \min\left\{ \frac{t^2}{nK^2}, \frac{t}{K} \right\} \right).$$

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

$$\|\Theta^\top Z_0\|_2 \leqslant O(\sqrt{n((d+m) + \log(1/\delta))}). \tag{2.20}$$

Combining equations (2.19) and (2.20) with a union bound, if $n \geqslant \Omega(d + m + \log(1/\delta))$, then

$$\|\mu - \hat{\mu}\|_2 \leqslant O\left(\sqrt{\frac{(d+m) + \log(1/\delta)}{n}} \right).$$

□

Proof of Lemma 2.5.9. Under the location-family parameterization, we can write

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

so the gradients are given by

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

This representation allows us to write

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

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

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

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

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

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

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

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

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

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

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

□

Proof of Lemma 2.5.10. By applying the location family parameterization as in the proof of Lemma 2.5.9, we get

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

Using the chain rule and the triangle inequality,

$$\begin{aligned}\|\nabla \mathrm{PR}(\theta) - \nabla \mathrm{PR}(\theta')\|_2 \leqslant \|&\mathop{\mathbb{E}}_{z_0 \sim \mathcal{D}_0} \nabla_\theta \ell(z_0 + \mu\theta; \theta) - \nabla_\theta \ell(z_0 + \mu\theta'; \theta')\|_2 \\ &+ \|\mathop{\mathbb{E}}_{z_0 \sim \mathcal{D}_0} \mu^\top \nabla_z \ell(z_0 + \mu\theta; \theta) - \mu^\top \nabla_z \ell(z_0 + \mu\theta'); \theta')\|_2. \end{aligned} \tag{2.21}$$

For the first term in equation (2.21), adding and subtracting $\nabla_\theta \ell(z + \mu\theta'; \theta)$ and using the triangle inequality gives

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

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 (2.21), similarly adding and subtracting $\mu^\top \nabla_z \ell(z + \mu\theta'; \theta)$ and using the triangle inequality gives

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

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. □

2.5.11 Proof of Lemma 2.3.1

Notice that $\mathrm{PR}(\theta) - \mathrm{PR}(\theta') = (\mathrm{DPR}(\theta, \theta) - \mathrm{DPR}(\theta, \theta')) + (\mathrm{DPR}(\theta, \theta') - \mathrm{DPR}(\theta', \theta'))$. We bound the first difference using Lipschitzness of ℓ in θ as $|\mathrm{DPR}(\theta, \theta) - \mathrm{DPR}(\theta, \theta')| = |\mathrm{E}_{z \sim \mathcal{D}(\theta)}[\ell(z; \theta) - \ell(z; \theta')]| \leqslant L_\theta\|\theta - \theta'\|$. For the second term we combine Definition 2.1.3 and Lipschitzness of ℓ in z via the Kantorovich-Rubinstein duality theorem. In particular, we get $|\mathrm{DPR}(\theta, \theta') - \mathrm{DPR}(\theta', \theta')| = |\mathrm{E}_{z \sim \mathcal{D}(\theta)}\ell(z; \theta') - \mathrm{E}_{z \sim \mathcal{D}(\theta')}\ell(z; \theta')| \leqslant \varepsilon L_z\|\theta - \theta'\|$. Putting both bounds together, we obtain the claimed Lipschitz bound.

2.5.12 Proof of Proposition 2.3.3

We construct a γ-cover of the parameter space, denoted $\mathcal{S}_\gamma$, and deploy all models in this cover. This gives us access to the distributions $\{\mathcal{D}(\theta) : \theta \in \mathcal{S}_\gamma\}$. Using this information, for any $\theta \in \Theta$ we can compute

$$\widehat{\mathrm{PR}}(\theta) = \mathrm{DPR}(\Pi_{\mathcal{S}_\gamma}(\theta), \theta) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}(\Pi_{\mathcal{S}_\gamma}(\theta))} \ell(z; \theta),$$

where $\Pi_{\mathcal{S}_\gamma}(\theta) := \arg\min_{\theta' \in \mathcal{S}_\gamma} \|\theta' - \theta\|$ is the projection onto $\mathcal{S}_\gamma$. Note that $\|\theta - \Pi_{\mathcal{S}_\gamma}(\theta)\| \leqslant \gamma$ all $\theta \in \Theta$ since $\mathcal{S}_\gamma$ is a cover. Therefore, for any $\theta \in \Theta$, we can bound $\mathrm{PR}(\theta)$ as

$$\begin{aligned} \mathrm{PR}(\theta) &\leqslant \mathrm{DPR}(\Pi_{\mathcal{S}_\gamma}(\theta), \theta) + L_z \varepsilon \|\Pi_{\mathcal{S}_\gamma}(\theta) - \theta\| \\ &\leqslant \mathrm{DPR}(\Pi_{\mathcal{S}_\gamma}(\theta), \theta) + L_z \varepsilon \gamma \\ &= \widehat{\mathrm{PR}}(\theta) + L_z \varepsilon \gamma. \end{aligned}$$

Similarly we obtain $\mathrm{PR}(\theta) \geqslant \widehat{\mathrm{PR}}(\theta) - L_z \varepsilon \gamma$, which completes the proof.

2.5.13 Proof of Proposition 2.3.4

We will show that $\mathrm{PR}_{\mathrm{LB}}(\theta) \leqslant \mathrm{PR}(\theta)$ and $\mathrm{PR}_{\min} \geqslant \mathrm{PR}(\theta_{\mathrm{PO}})$; these two facts immediately imply $\Delta(\theta) := \mathrm{PR}(\theta) - \mathrm{PR}(\theta_{\mathrm{PO}}) \geqslant \mathrm{PR}_{\mathrm{LB}}(\theta) - \mathrm{PR}_{\min}$.

The first bound follows because $\mathrm{PR}(\theta) = \mathrm{DPR}(\theta, \theta) \geqslant \mathrm{DPR}(\theta', \theta) - L_z \varepsilon \|\theta' - \theta\|$ for all θ', where we use $(L_z\varepsilon)$-Lipschitzness of DPR in the first argument. Similarly, the second bound follows because

$$\mathrm{PR}(\theta_{\mathrm{PO}}) = \min_\theta \mathrm{DPR}(\theta, \theta) \leqslant \min_\theta (\mathrm{DPR}(\theta', \theta) + L_z \varepsilon \|\theta - \theta'\|),$$

for all θ'.

2.5.14 Proof of Theorem 2.3.4

We prove a regret bound for Algorithm 2. We use $\mathrm{Reg}_{\mathrm{ph}}(p_1 : p_2)$ to denote the regret incurred from phase p_1 to phase p_2:

$$\mathrm{Reg}_{\mathrm{ph}}(p_1 : p_2) = \mathbb{E} \sum_{p=p_1}^{p_2} \Delta(\theta_p).$$

We let $\mathrm{Reg}_{\mathrm{ph}}(0 : p) \equiv \mathrm{Reg}_{\mathrm{ph}}(p)$. For phases p that happen after the time horizon T, we assume that the incurred regret is 0; for example, if phases $p_1 \leqslant p_2$ happen after T, then $\mathrm{Reg}_{\mathrm{ph}}(p_1 : p_2) = 0$.

Clean event

First, we define a clean event that guarantees that the estimates $\widehat{\mathrm{DPR}}(\theta, \theta')$ are close to the true values $\mathrm{DPR}(\theta, \theta')$ at all phases. The clean event essentially guarantees uniform convergence over $\widehat{\mathrm{DPR}}(\theta, \cdot)$ for every $\theta \in \mathcal{P}_p$.

Definition 2.5.1 (Clean event). *Denote the "clean event" by*

$$E_{\mathrm{clean}} = \left\{ \forall p : \sup_{\theta \in \mathcal{P}_p} \sup_{\theta' \in \Theta} \left| \widehat{\mathrm{DPR}}(\theta, \theta') - \mathrm{DPR}(\theta, \theta') \right| \leqslant \frac{2\mathfrak{C}^*(\ell) + 3\sqrt{\log(T)}}{\sqrt{n_p m_0}} \right\}, \tag{2.22}$$

where $\mathcal{P}_p$ is the set of all models deployed in phase p during time horizon T.

We show that the clean event occurs with high probability.

Lemma 2.5.11. *The clean event holds with high probability,*

$$\mathbb{P}\left\{E_{\mathrm{clean}}\right\} \geqslant 1 - T^{-2}.$$

Proof. We consider each interval of length n_p in phase p, during which the same model is deployed, separately, and then take a union bound over these intervals across all phases. Therefore, we will say interval s in phase p to refer to steps $(s-1)n_p + 1, \ldots, sn_p$ in phase p. For the sake of this proof, we consider a "counterfactual" set of samples for each model θ that augments the set of actually observed samples. In particular, for interval s in phase p, we let $\{z_j^{\theta,s}\}_{j=1}^{n_p m_0}$ denote i.i.d. samples from $\mathcal{D}(\theta)$. The samples for different time intervals and different phases are independent. When model θ is deployed, we observe the samples corresponding to the interval in which θ is deployed.

For each phase p and each time interval s within phase p, let $E_{\mathrm{end}}^{s,p}$ denote the event that phase p terminates strictly before interval s is reached. Let $E_{\mathrm{clean}}^{s,p}$ denote the event that one of the following two holds:

(E1) $E_{\mathrm{end}}^{s,p}$ occurs;

(E2) $E_{\mathrm{end}}^{s,p}$ does not occur, and for the model θ_s deployed in time interval s it holds that:

$$\sup_{\theta' \in \Theta} \left| \widehat{\mathrm{DPR}}(\theta_s, \theta') - \mathrm{DPR}(\theta_s, \theta') \right| \leqslant \frac{2\mathfrak{C} + 3\sqrt{\log(T)}}{\sqrt{n_p m_0}},$$

where θ_s is a random variable.

The probability that $E^{s,p}_{\text{clean}}$ does not occur is at most:

$$\begin{aligned}
&\mathbb{P}\left[\neg E^{s,p}_{\text{end}} \ \&\ \sup_{\theta'\in\Theta}\left|\widehat{\mathrm{DPR}}(\theta_s,\theta') - \mathrm{DPR}(\theta_s,\theta')\right| > \frac{2\mathfrak{C} + 3\sqrt{\log(T)}}{\sqrt{n_p m_0}}\right]\\
&= \mathbb{P}\left[\neg E^{s,p}_{\text{end}}\right]\cdot\mathbb{P}\left[\sup_{\theta'\in\Theta}\left|\widehat{\mathrm{DPR}}(\theta_s,\theta') - \mathrm{DPR}(\theta_s,\theta')\right| > \frac{2\mathfrak{C} + 3\sqrt{\log(T)}}{\sqrt{n_p m_0}}\middle|\neg E^{s,p}_{\text{end}}\right]\\
&\leqslant \mathbb{P}\left[\sup_{\theta'\in\Theta}\left|\widehat{\mathrm{DPR}}(\theta_s,\theta') - \mathrm{DPR}(\theta_s,\theta')\right| > \frac{2\mathfrak{C} + 3\sqrt{\log(T)}}{\sqrt{n_p m_0}}\middle|\neg E^{s,p}_{\text{end}}\right].
\end{aligned}$$

We can equivalently write this as

$$\mathbb{P}\left[\sup_{\theta'\in\Theta}\left|\frac{1}{n_p m_0}\sum_{j=1}^{n_p m_0}\ell(z_j^{\theta_s,s};\theta') - \mathrm{DPR}(\theta_s,\theta')\right| > \frac{2\mathfrak{C} + 3\sqrt{\log(T)}}{\sqrt{n_p m_0}}\middle|\neg E^{s,p}_{\text{end}}\right]$$

$$= \mathbb{E}_{\theta\sim\theta_s}\left[\mathbb{P}\left[\sup_{\theta'\in\Theta}\left|\frac{1}{n_p m_0}\sum_{j=1}^{n_p m_0}\ell(z_j^{\theta,s};\theta') - \mathrm{DPR}(\theta,\theta')\right| > \frac{2\mathfrak{C} + 3\sqrt{\log(T)}}{\sqrt{n_p m_0}}\middle|\neg E^{s,p}_{\text{end}}, \theta_s = \theta\right]\right].$$

To upper bound this expression, it suffices to show an upper bound on

$$\mathbb{P}\left[\sup_{\theta'\in\Theta}\left|\frac{1}{n_p m_0}\sum_{j=1}^{n_p m_0}\ell(z_j^{\theta,s};\theta') - \mathrm{DPR}(\theta,\theta')\right| > \frac{2\mathfrak{C} + 3\sqrt{\log(T)}}{\sqrt{n_p m_0}}\middle|\neg E^{s,p}_{\text{end}}, \theta_s = \theta\right]$$

that holds for every θ. The first observation is that for any θ, the samples $\{z_j^{\theta,s}\}_{j=1}^{n_p m_0}$ are independent of the event $\{\theta_s = \theta, \neg E^{s,p}_{\text{end}}\}$, since the event depends only on the samples collected in previous time intervals and phases. This means that the above probability is equal to:

$$\mathbb{P}\left[\sup_{\theta'\in\Theta}\left|\frac{1}{n_p m_0}\sum_{j=1}^{n_p m_0}\ell(z_j^{\theta,s};\theta') - \mathrm{DPR}(\theta,\theta')\right| > \frac{2\mathfrak{C} + 3\sqrt{\log(T)}}{\sqrt{n_p m_0}}\right].$$

Let ε_j denote i.i.d. Rademacher random variables. Then, we can observe that with probability $1 - T^{-3}$, it holds that:

$$\begin{aligned}
&\sup_{\theta' \in \Theta} \left| \frac{1}{n_p m_0} \sum_{j=1}^{n_p m_0} \ell(z_j^{\theta,s}; \theta') - \mathrm{DPR}(\theta, \theta') \right| \\
&\leqslant \mathbb{E}\left[\sup_{\theta' \in \Theta} \left| \frac{1}{n_p m_0} \sum_{j=1}^{n_p m_0} \ell(z_j^{\theta,s}; \theta') - \mathrm{DPR}(\theta, \theta') \right| \right] + \sqrt{\frac{6\log(T)}{n_p m_0}} \\
&\leqslant 2 \cdot \mathbb{E}\left[\sup_{\theta' \in \Theta} \left| \frac{1}{n_p m_0} \sum_{j=1}^{n_p m_0} \ell(z_j^{\theta,s}; \theta') \cdot \varepsilon_j \right| \right] + \sqrt{\frac{6\log(T)}{n_p m_0}} \\
&\leqslant \frac{2}{\sqrt{n_p m_0}} \cdot \sup_{n \geqslant 1} \sqrt{n}\, \mathbb{E}\left[\sup_{\theta' \in \Theta} \left| \frac{1}{n} \sum_{j=1}^{n} \ell(z_j^{\theta}; \theta') \cdot \varepsilon_j \right| \right] + \sqrt{\frac{6\log(T)}{n_p m_0}} \\
&\leqslant \frac{2\mathfrak{C}^*(\ell) + 3\sqrt{\log(T)}}{\sqrt{n_p m_0}},
\end{aligned}$$

where the first step follows from the bounded differences inequality and the second step follows from a classical symmetrization argument. In the penultimate step we let $\{z_j^{\theta}\}_{j \in \mathbb{N}}$ denote an infinite sequence of samples from $\mathcal{D}(\theta)$. Putting this all together, we have that:

$$1 - \mathbb{P}\left[E_{\text{clean}}^{s,p}\right] \leqslant T^{-3}.$$

Finally, using that there are at most T intervals before time horizon T (across all phases), by a union bound we see that:

$$1 - \mathbb{P}\left[E_{\text{clean}}\right] \leqslant T^{-2},$$

as desired.

□

Suboptimality of the active set

We show that the elimination strategy in Algorithm 2 will never eliminate any performatively optimal point.

Lemma 2.5.12. *On the clean event* (2.22)*, any performatively optimal point* $\theta_{\mathrm{PO}} \in \arg\min_\theta \mathrm{PR}(\theta)$ *will always remain in* $\mathcal{A}$*.*

Proof. It suffices to show that θ_{PO} cannot be eliminated in Step 14 of Algorithm 2. Fix any phase p and denote by $\mathcal{P}_p$ the running set of deployed points at any point during phase p.

Then, we have:

$$\begin{aligned}
\mathrm{PR}_{\mathrm{LB}}(\theta_{\mathrm{PO}}) &= \max_{\theta'\in\mathcal{P}_p}\left(\widehat{\mathrm{DPR}}(\theta',\theta_{\mathrm{PO}}) - L_z\varepsilon\|\theta_{\mathrm{PO}}-\theta'\|\right)\\
&\leqslant \max_{\theta'\in\mathcal{P}_p}\left(\mathrm{DPR}(\theta',\theta_{\mathrm{PO}}) - L_z\varepsilon\|\theta_{\mathrm{PO}}-\theta'\|\right) + \gamma_p\\
&\leqslant \mathrm{PR}(\theta_{\mathrm{PO}}) + \gamma_p\\
&= \min_{\theta}\mathrm{DPR}(\theta,\theta) + \gamma_p\\
&\leqslant \min_{\theta}\min_{\theta'\in\mathcal{P}_p}\mathrm{DPR}(\theta',\theta) + L_z\varepsilon\|\theta-\theta'\| + \gamma_p\\
&\leqslant \min_{\theta}\min_{\theta'\in\mathcal{P}_p}\widehat{\mathrm{DPR}}(\theta',\theta) + L_z\varepsilon\|\theta'-\theta\| + 2\gamma_p\\
&= \mathrm{PR}_{\min} + 2\gamma_p.
\end{aligned}$$

Therefore, $\mathrm{PR}_{\mathrm{LB}}(\theta_{\mathrm{PO}}) \leqslant \mathrm{PR}_{\min} + 2\gamma_p$, implying that θ_{PO} cannot be removed from $\mathcal{A}$ during phase p. Since this is true for any phase p, that completes the proof of the lemma. □

We next show that the elimination strategy is sufficiently effective that all models that remain active after a given phase p have suboptimality at most $8\gamma_p$.

Lemma 2.5.13. *On the clean event* (2.22)*, after phase p all models $\theta \in \mathcal{A}$ satisfy $\Delta(\theta) \leqslant 8\gamma_p$.*

Proof. Fix a phase p. We will analyze $\mathcal{P}_p$ at the *end* of phase p. The proof relies on two key facts:

(F1) If θ is active after phase p, then $\|\theta-\Pi_{\mathcal{P}_p}(\theta)\| \leqslant r_p$, where $\Pi_{\mathcal{P}_p}(\theta) = \arg\min_{\theta'\in\mathcal{P}_p}\|\theta-\theta'\|$.

(F2) θ_{PO} is active after phase p.

The first fact follows since during phase p net points cannot be eliminated from $\mathcal{S}_p$ in Step 13 while some parameter within an r_p-neighborhood is active. The second fact is proved in Lemma 2.5.12. Note that from fact (F1) it further follows that there is always a model in $\mathcal{P}_p$ within the r_p-neighborhood of θ_{PO}.

Now suppose that θ is active after phase p. Then, we have:

$$\begin{aligned}
\mathrm{PR}(\theta) &\leqslant \mathrm{DPR}(\Pi_{\mathcal{P}_p}(\theta),\theta) + L_z\varepsilon\|\Pi_{\mathcal{P}_p}(\theta)-\theta\|\\
&\leqslant \widehat{\mathrm{DPR}}(\Pi_{\mathcal{P}_p}(\theta),\theta) + L_z\varepsilon\|\Pi_{\mathcal{P}_p}(\theta)-\theta\| + \gamma_p\\
&\leqslant \min_{\theta'}\left(\widehat{\mathrm{DPR}}(\Pi_{\mathcal{P}_p}(\theta'),\theta') + L_z\varepsilon\|\Pi_{\mathcal{P}_p}(\theta')-\theta'\|\right) + 2L_z\varepsilon\|\Pi_{\mathcal{P}_p}(\theta)-\theta\| + 3\gamma_p,
\end{aligned}$$

where we used the definitions of $\mathrm{PR}_{\min}$ and $\mathrm{PR}_{\mathrm{LB}}(\theta)$, together with the fact that $\mathrm{PR}_{\mathrm{LB}}(\theta) \leqslant \mathrm{PR}_{\min} + 2\gamma_p$ for active models. Now choosing $\theta' = \theta_{\mathrm{PO}}$, applying (F1), (F2), and accounting for finite-sample uncertainty we find

$$\begin{aligned}\mathrm{PR}(\theta) &\leqslant \widehat{\mathrm{DPR}}(\Pi_{\mathcal{P}_p}(\theta_{\mathrm{PO}}), \theta_{\mathrm{PO}}) + L_z\varepsilon\|\Pi_{\mathcal{P}_p}(\theta_{\mathrm{PO}}) - \theta_{\mathrm{PO}}\| + 2L_z\varepsilon\|\Pi_{\mathcal{P}_p}(\theta) - \theta\| + 3\gamma_p \\ &\leqslant \widehat{\mathrm{DPR}}(\Pi_{\mathcal{P}_p}(\theta_{\mathrm{PO}}), \theta_{\mathrm{PO}}) + 3L_z\varepsilon r_p + 3\gamma_p \\ &\leqslant \mathrm{DPR}(\theta_{\mathrm{PO}}, \theta_{\mathrm{PO}}) + L_z\varepsilon\|\Pi_{\mathcal{P}_p}(\theta_{\mathrm{PO}}) - \theta_{\mathrm{PO}}\| + 3L_z\varepsilon r_p + 4\gamma_p \\ &\leqslant \mathrm{PR}(\theta_{\mathrm{PO}}) + 4(L_z\varepsilon r_p + \gamma_p) \\ &= \mathrm{PR}(\theta_{\mathrm{PO}}) + 8\gamma_p,\end{aligned}$$

where we use the fact that $r_p = \frac{\gamma_p}{L_z\varepsilon}$. Rearranging the terms we obtain $\Delta(\theta) = \mathrm{PR}(\theta) - \mathrm{PR}(\theta_{\mathrm{PO}}) \leqslant 8\gamma_p$ as claimed in Lemma 2.5.13. □

Bounding the number of suboptimal deployments

For $i \geqslant 1$, we consider the suboptimality bands

$$\mathcal{E}_i = \left\{\theta : \Delta(\theta) \in [8 \cdot 2^{-i} L_z\varepsilon, 16 \cdot 2^{-i} L_z\varepsilon)\right\}.$$

In the following lemma, we bound the number of times that models in $\mathcal{E}_i$ can be deployed in a given phase.

Lemma 2.5.14. *Suppose that the clean event (2.22) holds. For $i \geqslant 1$, in phase $\log_2(1/(L_z\varepsilon)) \leqslant p \leqslant \log_2(1/(L_z\varepsilon)) + i + 1$, the number of models in $\mathcal{E}_i$ that are deployed is at most $\mathcal{O}\left((3/r_p)^{d_0}\right)$ in expectation, where d_0 is the $(L_z\varepsilon)$-sequential zooming dimension.*

To provide intuition for Lemma 2.5.14, it is informative to consider a weaker version of the lemma where d_0 is taken to be the $(L_z\varepsilon)$-zooming dimension rather than the $(L_z\varepsilon)$-sequential zooming dimension. To see why this weaker version of the lemma is true, notice that at the beginning of phase p, the set of active models $\mathcal{A}$ is a subset of $\{\theta : \Delta(\theta) \leqslant 8\gamma_{p-1}\} = \{\theta : \Delta(\theta) \leqslant 16\gamma_p\} = \{\theta : \Delta(\theta) \leqslant 16 L_z\varepsilon r_p\}$. The set of models deployed in phase p is contained in a minimal r_p-net of $\mathcal{A}$. Notice that $r_p \geqslant 2^{-(i+1)}$. By the definition of zooming dimension, we know that at most a multiple of $\left(\frac{3}{r_p}\right)^{d_0}$ elements from the set $\{\theta : \Delta(\theta) \in [8 \cdot 2^{-i} L_z\varepsilon, 16 \cdot 2^{-i} L_z\varepsilon)\} = \left\{\theta : \Delta(\theta) \in [16 \cdot 2^{-(i+1)} L_z\varepsilon, 32 \cdot 2^{-(i+1)} L_z\varepsilon)\right\}$ are deployed, as desired.

The proof of Lemma 2.5.14 boils down to refining this proof sketch to account for the sequential elimination aspect of Algorithm 2.

Proof. For the purposes of this analysis, we condition on the clean event.

Fix a phase $\log_2(1/(L_z\varepsilon)) \leqslant p \leqslant \log_2(1/(L_z\varepsilon)) + i + 1$. Let $\mathcal{S}_p^0$ be the covering of $\mathcal{A}$ chosen at the beginning of phase p, and let π be an ordering of $\mathcal{S}_p^0$ chosen uniformly at random. It is not difficult to see that Algorithm 2 is equivalent to drawing π at the beginning of the phase, and deploying models in the order given by π (naturally, skipping those that get eliminated). For technical convenience, we analyze this reformulation of the algorithm.

Condition on a realization π, and let $\mathcal{P}_p \subseteq \mathcal{S}_p^0$ be the set of models that are ultimately get deployed. Note that $\mathcal{P}_p$ depends on the randomness arising from finite-sample noise at each step of the phase. We will show a bound on $|\mathcal{P}_p|$ that deterministically holds on the clean event. In particular, consider the models $\theta \in \mathcal{S}_p^0 \cap \{\theta : 8L_z\varepsilon r_i \leqslant \Delta(\theta) < 16L_z\varepsilon r_i\}$ such that:

$$\mathrm{PR}_{\mathrm{LB}}^{r_p}(\pi(\theta)) \leqslant \mathrm{PR}_{\min}(\pi(\theta)) + 4L_z\varepsilon r_p = \mathrm{PR}_{\min}(\pi(\theta)) + 4\gamma_p. \tag{2.23}$$

We will show that $\mathcal{P}_p$ is a subset of such models.

Suppose that $\theta_{\mathrm{net}} \in \mathcal{S}_p^0$ is deployed in phase p. Then, that means that there exists $\theta'' \in \mathrm{Ball}_{r_p}(\theta_{\mathrm{net}})$ that remains active after the first $\pi(\theta_{\mathrm{net}}) - 1$ deployments; that is:

$$\begin{aligned}\max_{\theta':\pi(\theta')<\pi(\theta_{\mathrm{net}})}(\widehat{\mathrm{DPR}}(\theta',\theta'') - L_z\varepsilon\|\theta'-\theta''\|) &= \mathrm{PR}_{\mathrm{LB}}(\theta'')\\ &\leqslant \mathrm{PR}_{\min} + 2\gamma_p\\ &= \min_{\theta}\min_{\theta':\pi(\theta')<\pi(\theta_{\mathrm{net}})}(\widehat{\mathrm{DPR}}(\theta',\theta) + L_z\varepsilon\|\theta'-\theta\|) + 2\gamma_p.\end{aligned}$$

Since the clean event holds, we know that:

$$\max_{\theta':\pi(\theta')<\pi(\theta_{\mathrm{net}})}(\mathrm{DPR}(\theta',\theta'') - L_z\varepsilon\|\theta'-\theta''\|) - \gamma_p \leqslant \min_{\theta}\min_{\theta':\pi(\theta')<\pi(\theta_{\mathrm{net}})}(\mathrm{DPR}(\theta',\theta) + L_z\varepsilon\|\theta'-\theta\|) + 3\gamma_p.$$

Rearranging, this means that:

$$\begin{aligned}&\max_{\theta':\pi(\theta')<\pi(\theta_{\mathrm{net}})}(\mathrm{DPR}(\theta',\theta'') - L_z\varepsilon\|\theta'-\theta''\|)\\ &\qquad \leqslant \min_{\theta}\min_{\theta':\pi(\theta')<\pi(\theta_{\mathrm{net}})}(\mathrm{DPR}(\theta',\theta) + L_z\varepsilon\|\theta'-\theta\|) + 4\gamma_p = \mathrm{PR}_{\min}(\pi(\theta_{\mathrm{net}})) + 4\gamma_p.\end{aligned}$$

This further implies that:

$$\mathrm{PR}_{\mathrm{LB}}^{r_p}(\pi(\theta_{\mathrm{net}})) = \min_{\theta''\in\mathrm{Ball}_{r_p}(\theta_{\mathrm{net}})}\max_{\theta':\pi(\theta')<\pi(\theta_{\mathrm{net}})}(\mathrm{DPR}(\theta',\theta'') - L_z\varepsilon\|\theta'-\theta''\|) \leqslant \mathrm{PR}_{\min}(\pi(\theta_{\mathrm{net}})) + 4\gamma_p.$$

We see that any $\theta_{\mathrm{net}} \in \mathcal{P}_p$ must satisfy condition (2.23). By the definition of sequential zooming dimension, we know that the expected number of models in $\mathcal{E}_i$ that satisfy (2.23), where the expectation is taken over the randomness of π, is at most a multiple of $\left(\frac{3}{r_p}\right)^{d_0}$, hence $\mathbb{E}\,|\mathcal{P}_p \cap \mathcal{E}_i| \leqslant \mathcal{O}\left(\left(\frac{3}{r_p}\right)^{d_0}\right)$, as desired. □

Regret bound on the clean event

To bound the regret on the clean event, we break the analysis into two cases: (a) the first $\log_2(1/(L_z\varepsilon))$ phases, and (b) all remaining phases.

Lemma 2.5.15. *Suppose that the clean event (2.22) holds. In the first $\lfloor \log_2(1/(L_z\varepsilon))\rfloor$ phases, the algorithm has incurred regret at most*

$$\mathrm{Reg}_{\mathrm{ph}}\left(\lfloor \log_2(1/(L_z\varepsilon))\rfloor\right) = \mathcal{O}\left(\sqrt{\frac{T}{m_0}}\left(\sqrt{\log T} + \mathfrak{C}\right)\right).$$

Proof. During phases $p \leqslant \log_2(1/(L_z\varepsilon))$, we deploy a single model since $r_p \geqslant 1$ and Θ is assumed to have radius 1.

We break the first $\lfloor \log_2(1/(L_z\varepsilon))\rfloor$ phases into two cases. For a value of $N \geqslant 0$ specified later, we consider cases $p < N$ and $p \geqslant N$ separately.

Case 1: phases $N \leqslant p \leqslant \lfloor \log_2(1/(L_z\varepsilon))\rfloor$**.** By Lemma 2.5.13, we see that the model deployed in phase N must have suboptimality at most $8\cdot 2^{-N+1} = 2^{-N+4}$. Since the algorithm runs for at most T time steps, this means that the total regret incurred in these phases is at most $T \cdot 2^{-N+4}$.

Case 2: phases $0 \leqslant p < \min\{N, \lfloor \log_2(1/(L_z\varepsilon))\rfloor\}$**.** By Lemma 2.5.13, we know that the model deployed in phase p must have suboptimality at most $8 \cdot 2^{-p+1} = 2^{-p+4}$. Moreover, this model is deployed for $n_p = \left\lceil \frac{\left(2\mathfrak{C}+3\sqrt{\log T}\right)^2}{\gamma_p^2 m_0} \right\rceil$ steps. The regret incurred up to phase N can thus be bounded as:

$$\begin{aligned}\mathrm{Reg}_{\mathrm{ph}}(N) &\leqslant \sum_{p=0}^{N-1} n_p 2^{-p+4}\\ &\leqslant 16\sum_{p=0}^{N-1} 2^{-p}\left\lceil \frac{2^{2p}(2\mathfrak{C}+3\sqrt{\log T})^2}{m_0}\right\rceil.\end{aligned}$$

Since we assume $m_0 = o((\mathfrak{C}+\sqrt{\log T})^2)$, for a large enough T we have $n_p \geqslant 1$ and thus $\lceil n_p \rceil \leqslant 2n_p$. Therefore,

$$\begin{aligned}\mathrm{Reg}_{\mathrm{ph}}(N) &\leqslant C\sum_{p=0}^{N-1} 2^{-p}\frac{2^{2p}(2\mathfrak{C}+3\sqrt{\log T})^2}{m_0}\\ &\leqslant C\frac{(2\mathfrak{C}+3\sqrt{\log T})^2}{m_0}\left(\sum_{p=0}^{N-1} 2^p\right)\\ &\leqslant C\cdot 2^N\frac{(2\mathfrak{C}+3\sqrt{\log T})^2}{m_0},\end{aligned}$$

for some large enough constant $C > 0$.

Putting the two cases together, on the clean event, the total regret incurred in phases $p = 0, \ldots, \lfloor \log_2(1/(L_z\varepsilon)) \rfloor$ can be upper bounded by

$$C \cdot 2^N \frac{(2\mathfrak{C} + 3\sqrt{\log T})^2}{m_0} + T \cdot 2^{-N+4}.$$

We can also trivially upper bound the regret by T, using the fact that the loss incurred at each step is at most 1. This means that we obtain a regret bound of:

$$\mathcal{O}\left(\min\left\{T, 2^N \frac{(2\mathfrak{C} + 3\sqrt{\log T})^2}{m_0} + T \cdot 2^{-N+4}\right\}\right).$$

We now choose N to minimize this bound. We let $\eta = 2^{-N}$ and optimize over $\eta \in (0, 1)$. Optimizing over η instead of an integral value of N changes the bound by constant factors at most. This means that we can upper bound the regret by:

$$\mathcal{O}\left(\min_{0<\eta\leqslant 1} \min\left\{T, \eta^{-1} \frac{\left(2\mathfrak{C} + 3\sqrt{\log T}\right)^2}{m_0} + T\eta\right\}\right).$$

If $\eta > 1$, then the minimum of the two terms would be T, which is at least as big as the above expression. Therefore, we can upper bound the above expression by:

$$\mathcal{O}\left(\min_{\eta>0}\left(\eta^{-1} \frac{\left(2\mathfrak{C} + 3\sqrt{\log T}\right)^2}{m_0} + T\eta\right)\right).$$

We set $\eta = \frac{3\sqrt{\log T} + 2\mathfrak{C}}{\sqrt{m_0 T}}$ and obtain a regret bound of:

$$\mathrm{Reg}_{\mathrm{ph}}\left(\lfloor \log_2(1/(L_z\varepsilon)) \rfloor\right) = \mathcal{O}\left(\sqrt{\frac{T}{m_0}}\left(\sqrt{\log T} + \mathfrak{C}\right)\right),$$

as desired. □

Lemma 2.5.16. *Suppose that the clean event* (2.22) *holds. Let $d \geqslant 0$ be such that for every $i \geqslant 0$ and every phase $p \in [\log_2(1/(L_z\varepsilon)), \log_2(1/(L_z\varepsilon)) + i + 1]$, the number of models in $\mathcal{E}_i = \{\theta : \Delta(\theta) \in [2^{-i+3}L_z\varepsilon, 2^{-i+4}L_z\varepsilon]\}$ that are deployed in phase p is upper bounded by $\mathcal{O}\left(\left(\frac{3}{r_p}\right)^d\right)$ in expectation. Then, the regret incurred in phases $p \geqslant \log_2(1/(L_z\varepsilon))$, within time horizon T, can be upper bounded as*

$$\mathrm{Reg}_{\mathrm{ph}}\left(\lceil \log_2(1/(L_z\varepsilon)) \rceil : \infty\right) \leqslant \mathcal{O}\left(T^{\frac{d+1}{d+2}} (L_z\varepsilon)^{\frac{d}{d+2}} \left(\frac{(\sqrt{\log T} + \mathfrak{C})^2}{m_0}\right)^{\frac{1}{d+2}}\right).$$

Proof. By Lemma 2.5.13, we see that all models θ that are active in phase $p = \lceil \log_2(1/L_z\varepsilon))\rceil$ or later have $\Delta(\theta) \leqslant 8L_z\varepsilon r_p \leqslant 8L_z\varepsilon$. We split these models into suboptimality bands and define, for each $i \geqslant 1$, the set:

$$\mathcal{E}_i = \left\{\theta : \Delta(\theta) \in [8 \cdot 2^{-i} L_z\varepsilon, 16 \cdot 2^{-i} L_z\varepsilon)\right\}.$$

Note that all models deployed starting with phase $\lceil \log_2(1/(L_z\varepsilon))\rceil$ are in $\cup_{i\geqslant 1}\mathcal{E}_i$. For a value of N specified later, we break the analysis into two cases.

Case 1: models in $\cup_{i>N}\mathcal{E}_i$. Since the algorithm runs for at most T time steps, the total regret incurred due to deploying models in $\cup_{i>N}\mathcal{E}_i$ is at most

$$T \cdot 16 \cdot 2^{-N-1} L_z\varepsilon \leqslant 8T2^{-N} L_z\varepsilon.$$

Case 2: models in $\cup_{1\leqslant i\leqslant N}\mathcal{E}_i$. By Lemma 2.5.13, we know that all models θ that are active is phases $p \geqslant N + \log_2(1/L_z\varepsilon)$ have $\Delta(\theta) \leqslant 82^{-p} = 8 \cdot 2^{-N} L_z\varepsilon = 16 \cdot 2^{-N-1} L_z\varepsilon$. This means that all models that are active after phase $N + \log_2(1/L_z\varepsilon)$ are in $\cup_{i>N}\mathcal{E}_i$. Thus, to bound the regret incurred by deploying models in $\cup_{1\leqslant i\leqslant N}\mathcal{E}_i$ in phase $\lceil \log_2(1/L_z\varepsilon)\rceil$ or later, we only need to consider phases $p = \lceil \log_2(1/L_z\varepsilon)\rceil, \ldots, N + \log_2(1/L_z\varepsilon)$.

For $1 \leqslant i \leqslant N$, consider $\mathcal{E}_i$. By Lemma 2.5.13, we know that any $\theta \in \mathcal{E}_i$ can only be active during phases $p \leqslant \log_2(1/L_z\varepsilon) + i + 1$. By assumption, in phase p, the number of points in $\mathcal{E}_i$ that are deployed is at most of the order $\left(\frac{3}{r_p}\right)^d$ in expectation. Moreover, each point is deployed n_p times. Putting this all together, the expected number of points in $\mathcal{E}_i$ deployed in phase p is at most:

$$\mathcal{O}\left(\left(\frac{3}{r_p}\right)^d n_p\right) = \mathcal{O}\left(\left(\frac{3}{r_p}\right)^d \frac{(2\mathfrak{C} + 3\sqrt{\log T})^2}{L_z^2\varepsilon^2 r_p^2 m_0}\right),$$

where we use the fact that, given the condition $m_0 = o((\mathfrak{C} + \sqrt{\log T})^2)$, $n_p \geqslant 1$ for large enough T and hence we can bound $\lceil n_p\rceil \leqslant 2n_p$. Take $p = j + \log_2(1/L_z\varepsilon))$; then, $r_p = 2^{-j}$. We sum over phases $\log_2(1/(L_z\varepsilon)) \leqslant p \leqslant \log_2(1/(L_z\varepsilon)) + i + 1$ to obtain that in expectation, the total number of times that these models are deployed is at most:

$$\mathcal{O}\left(\frac{3^d(2\mathfrak{C} + 3\sqrt{\log T})^2}{L_z^2\varepsilon^2 m_0}\sum_{j=0}^{i+1} 2^{j(d+2)}\right) = \mathcal{O}\left(\frac{3^d(2\mathfrak{C} + 3\sqrt{\log T})^2}{L_z^2\varepsilon^2 m_0} 2^{(i+1)(d+2)}\right).$$

Using the fact that the models have suboptimality at most $16 \cdot 2^{-i} L_z\varepsilon = 32 \cdot 2^{-(i+1)} L_z\varepsilon$, we see that the regret incurred by deploying models in $\mathcal{E}_i$ is upper bounded by:

$$\mathcal{O}\left(\frac{3^d(2\mathfrak{C} + 3\sqrt{\log T})^2}{L_z\varepsilon m_0} 2^{(i+1)(d+1)}\right).$$

We sum over $1 \leqslant i \leqslant N$ to obtain the total regret incurred due to deploying models in $\cup_{1\leqslant i\leqslant N}\mathcal{E}_i$:

$$\mathcal{O}\left(\frac{3^d(2\mathfrak{C}+3\sqrt{\log T})^2}{L_z\varepsilon m_0}2^{(N+2)(d+1)}\right).$$

Putting together the two cases we obtain a total regret bound of

$$\mathcal{O}\left(\frac{3^d(2\mathfrak{C}+3\sqrt{\log T})^2}{L_z\varepsilon m_0}2^{(N+2)(d+1)}+T2^{-N}L_z\varepsilon\right).$$

We also can upper bound the regret by $8TL_z\varepsilon$, since all models active after phase phase $\lfloor\log_2(1/(L_z\varepsilon))\rfloor$ have $\Delta(\theta)\leqslant 8L_z\varepsilon$ and there are at most T time steps in total. This means that we can bound the regret by:

$$\mathcal{O}\left(\min\left\{TL_z\varepsilon, \frac{3^d(2\mathfrak{C}+3\sqrt{\log T})^2}{L_z\varepsilon m_0}2^{(N+2)(d+1)}+T2^{-N}L_z\varepsilon\right\}\right).$$

We now choose N to minimize this bound. We let $\eta = 2^{-N}$ and choose some $\eta \in (0,1)$. The error from optimizing over $\eta \in (0,1)$ instead of an integral value of N contributes at most constant factors. This means that we can upper bound the regret by:

$$\mathcal{O}\left(\min\left\{TL_z\varepsilon, \frac{12^d(2\mathfrak{C}+3\sqrt{\log T})^2}{L_z\varepsilon m_0}\eta^{-(d+1)}+T\eta L_z\varepsilon\right\}\right),$$

for any $\eta \in (0,1)$. Note that, if $\eta \geqslant 1$, the second term in the bound is at least as large as the first term, hence we can choose any $\eta > 0$. In particular, we can further upper bound the regret by

$$\mathcal{O}\left(\min_{\eta>0}\left(\frac{12^d(2\mathfrak{C}+3\sqrt{\log T})^2}{L_z\varepsilon m_0}\eta^{-(d+1)}+T\eta L_z\varepsilon\right)\right).$$

Now, we set

$$\eta = \left(\frac{12^d\left(3\sqrt{\log T}+2\mathfrak{C}\right)^2}{TL_z^2\varepsilon^2 m_0}\right)^{\frac{1}{d+2}}.$$

Thus, we finally get a regret bound of

$$\mathcal{O}\left(T^{\frac{d+1}{d+2}}(L_z\varepsilon)^{\frac{d}{d+2}}\left(\frac{\left(\sqrt{\log T}+\mathfrak{C}\right)^2}{m_0}\right)^{\frac{1}{d+2}}\right),$$

as desired. □

Putting everything together

Now, we are ready to prove Theorem 2.3.4.

First, we handle the case where the clean event defined in (2.22) does not hold and the concentration bound is violated. By Lemma 2.5.11, this happens with probability at most T^{-2}. The regret incurred in each deployment is at most 1 and there are T deployments, so these events contribute a negligible factor T^{-1} to the expected regret.

For the case where the clean event holds we can build on Lemma 2.5.14, Lemma 2.5.15, and Lemma 2.5.16. From Lemma 2.5.15, we obtain a bound for the total regret incurred in phases up to $\lfloor \log_2(1/(L_z\varepsilon))\rfloor$. By Lemma 2.5.14 we can set the parameter d in Lemma 2.5.16 to be the $(L_z\varepsilon)$-sequential zooming dimension, and thus from Lemma 2.5.16 we obtain a regret bound for all later phases.

Putting all this together yields the desired bound.

2.5.15 Proof of Theorem 2.3.5

The proof of Theorem 2.3.5 relies on two key lemmas. One proves that $\mathcal{C}_t$ are valid confidence sets for μ_* at every step, and the other one proves a regret bound assuming that $\mathcal{C}_t$ are valid confidence sets.

Throughout we denote by $\mathcal{B}_m$ the unit ball in $\mathbb{R}^m$. For a vector x and matrix M, we will use the notation $\|x\|_M = \sqrt{x^\top M x}$.

An important object in the proofs will be $S_t := \sum_{i=1}^t \theta_i \bar{z}_{0,i}^\top$, where $\bar{z}_{0,i} = \frac{1}{m_0}\sum_{j=1}^{m_0} z_i^{(j)} - \mu_*^\top \theta_i$. Essentially $\bar{z}_{0,i}$ is the average over m_0 samples from $\mathcal{D}_0$, collected at step i. We will also denote $V_t(\lambda) = (\lambda I + \sum_{i=1}^t \theta_i\theta_i^\top)$, for an arbitrary offset $\lambda > 0$, and $V_t \equiv V_t(0)$. Note that in the algorithm statement we use $\Sigma_t = V_t\left(\frac{1}{m_0}\right)$.

Clean event

As for Algorithm 2, we introduce a clean event. In this case, the clean event will be defined as

$$E_{\text{clean}} = \{\forall t \in \mathbb{N} : \mu_* \in \mathcal{C}_t\}, \tag{2.24}$$

where $\mathcal{C}_t$ are the confidence sets constructed in Algorithm 3.

The technical subtlety lies in the fact that the points θ_t are chosen adaptively, hence one cannot simply apply standard least-squares confidence intervals to argue that the sets $\mathcal{C}_t$ are valid. The same difficulty is resolved in the analysis of the LinUCB algorithm and our proof builds on the proof technique of that analysis.

Before stating the main technical lemma, we start with an auxiliary result that we will use in the proof.

Lemma 2.5.17. *Suppose that $\mathcal{D}_0$ is 1-subgaussian. Then, for all $x \in \mathcal{B}_m$ and $y \in \mathbb{R}^{d_\Theta}$, the process*

$$M_t(x, y) = \exp\left(y^\top S_t x - \frac{1}{2m_0}\|y\|_{V_t}^2\right)$$

is a supermartingale with respect to the natural filtration, with $M_0(x, y) = 1$.

Proof. Since $\bar{z}_{0,i}$ are $\frac{1}{\sqrt{m_0}}$-subgaussian, we know that all one-dimensional projections are also $\frac{1}{\sqrt{m_0}}$-subgaussian, hence $\bar{z}_{0,i}^\top x$ are independent $\frac{1}{\sqrt{m_0}}$-subgaussian as well. Using this, we know

$$\mathbb{E}\left[\exp(y^\top \theta_t z_{0,t}^\top x) \;\middle|\; \mathcal{F}_{t-1}\right] \leqslant \exp\left(\frac{(y^\top \theta_t)^2}{2m_0}\right) = \exp\left(\frac{\|y\|_{\theta_t\theta_t^\top}^2}{2m_0}\right)$$

almost surely. Hence,

$$\begin{aligned}\mathbb{E}[M_t(x, y) \mid \mathcal{F}_{t-1}] &= \mathbb{E}\left[\exp\left(y^\top S_t x - \frac{1}{2m_0}\|y\|_{V_t}^2\right) \;\middle|\; \mathcal{F}_{t-1}\right]\\ &= M_{t-1}(x, y)\mathbb{E}\left[\exp\left(y^\top \theta_t z_{0,t}^\top x - \frac{1}{2m_0}\|y\|_{\theta_t\theta_t^\top}^2\right) \;\middle|\; \mathcal{F}_{t-1}\right]\\ &\leqslant M_{t-1}(x, y)\end{aligned}$$

almost surely. Furthermore, $M_0(x, y) = 1$ is trivially true. □

Now we are ready to state the main technical lemma about the validity of $\mathcal{C}_t$.

Lemma 2.5.18. *We have that*

$$\mathbb{P}\left\{E_{\text{clean}}\right\} \geqslant 1 - T^{-2}.$$

Proof. First we will show that for any $\delta \in (0, 1)$,

$$\mathbb{P}\left\{\exists t \in \mathbb{N} : \|V_t(\lambda)^{-1/2} S_t\|^2 \geqslant \frac{1}{m_0}\left(8m + 4\log\left(\frac{1}{\delta}\right) + 2\log\left(\frac{\det(V_t(\lambda))}{\lambda^{d_\Theta}}\right)\right)\right\} \leqslant \delta, \quad (2.25)$$

for all $\lambda > 0$.

Let $\Sigma = \frac{m_0}{\lambda} I \in \mathbb{R}^{d_\Theta \times d_\Theta}$ and let h be the density of $\mathcal{N}(0, \Sigma)$. Then, for any fixed $x \in \mathcal{B}_m$ and $M_t(x, y)$ as in Lemma 2.5.17, define

$$\bar{M}_t(x) = \int_{\mathbb{R}^{d_\Theta}} M_t(x, y)h(y) = \frac{1}{\sqrt{(2\pi)^{d_\Theta}\det(\Sigma)}}\int_{\mathbb{R}^{d_\Theta}} \exp\left(y^\top S_t x - \frac{1}{2m_0}\|y\|_{V_t}^2 - \frac{1}{2}\|y\|_{\Sigma^{-1}}^2\right) dy.$$

Notice that we can write

$$y^\top S_t x - \frac{1}{2m_0}\|y\|_{V_t}^2 - \frac{1}{2}\|y\|_{\Sigma^{-1}}^2 = \frac{1}{2}\|S_t x\|_{(\Sigma^{-1} + \frac{V_t}{m_0})^{-1}}^2 - \frac{1}{2}\left\|y - \left(\Sigma^{-1} + \frac{V_t}{m_0}\right)^{-1} S_t x\right\|_{\Sigma^{-1} + \frac{V_t}{m_0}}^2.$$

Thus, by integrating out the Gaussian density, we get

$$\begin{aligned}
&\bar{M_t}(x)\\
&= \exp\left(\frac{1}{2}\|S_t x\|^2_{(\Sigma^{-1}+\frac{V_t}{m_0})^{-1}}\right)\frac{1}{\sqrt{(2\pi)^{d_\Theta}\det(\Sigma)}}\int_{\mathbb{R}^{d_\Theta}}\exp\left(-\frac{1}{2}\left\|y-\left(\Sigma^{-1}+\frac{V_t}{m_0}\right)^{-1}S_t x\right\|^2_{\Sigma^{-1}+\frac{V_t}{m_0}}\right)dy\\
&= \exp\left(\frac{1}{2}\|S_t x\|^2_{(\Sigma^{-1}+\frac{V_t}{m_0})^{-1}}\right)\left(\frac{\det((\Sigma^{-1}+\frac{V_t}{m_0})^{-1})}{\det(\Sigma)}\right)^{1/2}\\
&= \exp\left(\frac{m_0}{2}\|V_t^{-1/2}(\lambda)S_t x\|^2\right)\left(\frac{\lambda^{d_\Theta}}{\det(V_t(\lambda))}\right)^{1/2}.
\end{aligned}$$

Now, by Lemma 20.3 in [105], since $M_t(x,y)$ is a supermartingale then $\bar{M_t}(x)$ is a non-negative supermartingale with $\bar{M_0}(x) = 1$. Thus, we can apply the maximal inequality to get

$$\begin{aligned}
&\mathbb{P}\left\{\exists t\in\mathbb{N} : \log\bar{M_t}(x) \geqslant \log(1/\delta)\right\}\\
&\quad = \mathbb{P}\left\{\exists t\in\mathbb{N} : \frac{m_0}{2}\|V_t(\lambda)^{-1/2}S_t x\|^2 - \frac{1}{2}\log\left(\frac{\det(V_t(\lambda))}{\lambda^{d_\Theta}}\right)\geqslant\log(1/\delta)\right\}\leqslant\delta. \qquad (2.26)
\end{aligned}$$

Inequality (2.26) is valid for all fixed $x\in\mathcal{B}_m$; to prove inequality (2.25), we use a covering argument. Let $N_{\frac{1}{2},m}$ denote a $\frac{1}{2}$-net of $\mathcal{B}_m$, and note that we can make $|N_{\frac{1}{2},m}|\leqslant 5^m$. Then,

$$\|V_t(\lambda)^{-1/2}S_t\| = \max_{x\in\mathcal{B}_m}\|V_t(\lambda)^{-1/2}S_t x\| \leqslant 2\max_{x\in N_{\frac{1}{2},m}}\|V_t(\lambda)^{-1/2}S_t x\|.$$

Therefore, we can apply a union bound to conclude that for all $s>0$,

$$\begin{aligned}
\mathbb{P}\left\{\exists t\in\mathbb{N} : \|V_t(\lambda)^{-1/2}S_t\|^2\geqslant s\right\} &\leqslant \mathbb{P}\left\{\exists t\in\mathbb{N} : \max_{x\in N_{1/2,m}}\|V_t(\lambda)^{-1/2}S_t x\|_2^2\geqslant\frac{s}{4}\right\}\\
&\leqslant\sum_{x\in N_{1/2,m}}\mathbb{P}\left\{\exists t\in\mathbb{N} : \|V_t(\lambda)^{-1/2}S_t x\|_2^2\geqslant\frac{s}{4}\right\}.
\end{aligned}$$

By picking $s=\frac{1}{m_0}(8m+4\log\frac{1}{\delta}+2\log(\frac{\det(V_t(\lambda))}{\lambda^{d_\Theta}}))\geqslant\frac{1}{m_0}(4\log\frac{5^m}{\delta}+2\log(\frac{\det(V_t(\lambda))}{\lambda^{d_\Theta}}))$ and applying Equation (2.26), we get

$$\mathbb{P}\left\{\exists t\in\mathbb{N} : \|V_t(\lambda)^{-1/2}S_t\|^2\geqslant\frac{1}{m_0}\left(8m+4\log\left(\frac{1}{\delta}\right)+2\log\left(\frac{\det(V_t(\lambda))}{\lambda^{d_\Theta}}\right)\right)\right\}\leqslant\sum_{x\in N_{1/2,m}}\frac{\delta}{5^m}\leqslant\delta.$$

This completes the proof of inequality (2.25).

It remains to relate this bound to the definition of $\mathcal{C}_t$. We can write

$$\hat{\mu}_t-\mu_* = V_t(\lambda)^{-1}S_t + V_t(\lambda)^{-1}V_t\mu_* - \mu_*,$$

and therefore

$$\begin{aligned}\|V_t(\lambda)^{1/2}(\hat{\mu}_t - \mu_*)\| &= \|V_t(\lambda)^{-1/2}S_t + V_t(\lambda)^{1/2}(V_t(\lambda)^{-1}V_t - I)\mu_*\| \\ &\leqslant \|V_t(\lambda)^{-1/2}S_t\| + \sqrt{\|\mu_*^\top(V_t(\lambda)^{-1}V_t - I)V_t(\lambda)(V_t(\lambda)^{-1}V_t - I)\mu_*\|} \\ &= \|V_t(\lambda)^{-1/2}S_t\| + \sqrt{\lambda}\sqrt{\|\mu_*^\top(I - V_t(\lambda)^{-1}V_t)\mu_*\|} \\ &= \|V_t(\lambda)^{-1/2}S_t\| + \sqrt{\lambda}\|\mu_*\|,\end{aligned}$$

where the second equality follows by writing $V_t = V_t(\lambda) - \lambda I$. Note additionally that by $\max\{\|\theta\| : \theta \in \Theta\} \leqslant 1$ and the AM-GM inequality,

$$\det(V_t(\lambda)) \leqslant \left(\frac{1}{d_\Theta}\text{trace}V_t(\lambda)\right)^{d_\Theta} \leqslant \left(\frac{d_\Theta\lambda + t}{d_\Theta}\right)^{d_\Theta}.$$

Applying Equation (2.25), setting $\delta = \frac{1}{T^2}$ and $\lambda = \frac{1}{m_0}$ completes the proof. □

Regret bound on the clean event

The place where the structure of the performative risk comes into play is the following lemma, where we relate the suboptimality of the deployed model θ_t to properties of the confidence set $\mathcal{C}_t$.

Lemma 2.5.19. *Suppose that the clean event* (2.24) *holds. Then, we can bound the suboptimality of θ_t by*

$$\Delta(\theta_t) \leqslant \min\left\{1, L_z \sup_{\mu,\mu' \in \mathcal{C}_t} \|(\mu - \mu')^\top\theta_t\|\right\}.$$

Proof. In what follows, all expectations are taken only over a sample $z_0 \sim \mathcal{D}_0$ independent of everything else (i.e., all other random quantities are conditioned on).

Since the loss is bounded, we know $\Delta(\theta_t) \leqslant 1$. For the other bound, notice that

$$\Delta(\theta_t) = \mathbb{E}\ell(z_0 + \mu_*^\top\theta_t; \theta_t) - \mathbb{E}\ell(z_0 + \mu_*^\top\theta_{\text{PO}}; \theta_{\text{PO}}).$$

By the definition of the algorithm and the clean event, we can lower bound the second term $\mathbb{E}\ell(z_0 + \mu_*^\top\theta_{\text{PO}}; \theta_{\text{PO}})$ as follows:

$$\mathbb{E}\ell(z_0 + \mu_*^\top\theta_{\text{PO}}; \theta_{\text{PO}}) \geqslant \text{PR}_{\text{LB}}(\theta_{\text{PO}}) \geqslant \text{PR}_{\text{LB}}(\theta_t) = \mathbb{E}\ell(z_0 + \tilde{\mu}_t^\top\theta_t; \theta_t),$$

for some $\tilde{\mu}_t \in \mathcal{C}_t$. This means that:

$$\Delta(\theta_t) \leqslant \mathbb{E}\ell(z_0 + \mu_*^\top\theta_t; \theta_t) - \mathbb{E}\ell(z_0 + \tilde{\mu}_t^\top\theta_t; \theta_t).$$

To finish, we use Lipschitzness of the loss to upper bound this by $L_z\|(\mu_* - \tilde{\mu}_t)^\top\theta_t\|$. Using the clean event, we can further upper bound this by $L_z \sup_{\mu,\mu' \in \mathcal{C}_t} \|(\mu - \mu')^\top\theta_t\|$ as desired. □

We now use this bound on the suboptimality of deployed models, along with the structure of the confidence sets, to bound the regret on the clean event.

Lemma 2.5.20. *Let $1 \leqslant \beta_1 \leqslant \beta_2 \leqslant \dots \beta_T$ and assume that the loss $\ell(z;\theta)$ is L_z-Lipschitz in z. Assume that the event*

$$\mu_* \in \mathcal{C}_t \subseteq \left\{ \mu \in \mathbb{R}^{d_\Theta \times m} : \left\| V_{t-1}^{1/2}\left(\frac{1}{m_0}\right)(\mu - \hat{\mu}_{t-1}) \right\|^2 \leqslant \beta_t \right\}$$

holds true, for all $2 \leqslant t \leqslant T$. Then, on this event, Algorithm 3 satisfies:

$$\sum_{t=1}^{T} \Delta(\theta_t) = \tilde{\mathcal{O}}\left(1 + \sqrt{d_\Theta T \beta_T \log\left(\frac{d_\Theta + T m_0}{d_\Theta}\right)} \max\{L_z, 1\}\right).$$

Proof. As in the proof of Lemma 2.5.19, all expectations are taken only over a sample $z_0 \sim \mathcal{D}_0$ independent of everything else (i.e., all other random quantities are conditioned on).

First, we separately bound the regret of the first step as $\mathcal{O}(1)$, using the fact that the loss is bounded in $[0,1]$.

For the remainder of the steps, we apply Lemma 2.5.19 to upper bound $\Delta(\theta_t)$. Using this, coupled with structure of $\mathcal{C}_t$, we can obtain the following upper bound, for any $\lambda > 0$:

$$\begin{aligned}
\Delta(\theta_t) &\leqslant \min\left\{1, L_z \sup_{\mu,\mu' \in \mathcal{C}_t} \|(\mu - \mu')^\top \theta_t\|\right\} \\
&\leqslant \min\left\{1, L_z \sup_{\mu,\mu' \in \mathcal{C}_t} \|(\mu - \mu')^\top V_{t-1}^{1/2}(\lambda)\| \cdot \|V_{t-1}^{-1/2}(\lambda)\theta_t\|\right\} \\
&\leqslant \min\left\{1, 2L_z \sqrt{\beta_t} \|V_{t-1}^{-1/2}(\lambda)\theta_t\|\right\} \\
&\leqslant 2\sqrt{\beta_T} \min\left\{1, L_z \|V_{t-1}^{-1/2}(\lambda)\theta_t\|\right\},
\end{aligned}$$

where the last line uses the fact that $\beta_T \geqslant \max\{1, \beta_t\}$.

By the Cauchy-Schwarz inequality,

$$\begin{aligned}\sum_{t=2}^{T}\Delta(\theta_t) &\leqslant \sqrt{T\sum_{t=2}^{T}\Delta(\theta_t)^2}\\ &\leqslant 2\sqrt{T\beta_T\sum_{t=2}^{T}\min\left\{1, L_z^2\|V_{t-1}^{-1/2}(\lambda)\theta_t\|^2\right\}}\\ &\leqslant 2\sqrt{T\beta_T\sum_{t=2}^{T}\min\left\{1, \max\{1, L_z^2\}\|V_{t-1}^{-1/2}(\lambda)\theta_t\|^2\right\}}\\ &\leqslant 2\sqrt{T\max\{1, L_z^2\}\beta_T\sum_{t=2}^{T}\min\left\{1, \|V_{t-1}^{-1/2}(\lambda)\theta_t\|^2\right\}}\\ &= 2\max\{1, L_z\}\sqrt{T\beta_T\sum_{t=2}^{T}\min\left\{1, \|V_{t-1}^{-1/2}(\lambda)\theta_t\|^2\right\}}.\end{aligned}$$

Finally, we use Lemma 19.4 in [105] that says

$$\sum_{t=2}^{T}\min\left\{1, \|V_{t-1}^{-1/2}(\lambda)\theta_t\|^2\right\} \leqslant 2d_\Theta\log\left(\frac{\text{trace}V_0(\lambda)+T}{d_\Theta\text{det}(V_0(\lambda))^{1/d_\Theta}}\right) = 2d_\Theta\log\left(\frac{d_\Theta\lambda+T}{d_\Theta\lambda}\right).$$

Using this expression in the equation above and setting $\lambda = \frac{1}{m_0}$ yields the final result. □

Putting everything together

We take $\sqrt{\beta_t} = \max\left\{1, \sqrt{\frac{1}{m_0}}M_* + \sqrt{\frac{8m+8\log T+2d_\Theta\log\left(\frac{d_\Theta+tm_0}{d_\Theta}\right)}{m_0}}\right\}$. By the constraint that $m_0 = o(\log T)$, we see that second branch dominates over the first one and so, for large enough T, $\sqrt{\beta_t} = \sqrt{\frac{1}{m_0}}M_* + \sqrt{\frac{8m+8\log T+2d_\Theta\log\left(\frac{d_\Theta+tm_0}{d_\Theta}\right)}{m_0}}$. Lemma 2.5.18 shows that:

$$\mu_* \in \mathcal{C}_t \subseteq \left\{\mu \in \mathbb{R}^{d_\Theta\times m} : \left\|V_{t-1}^{1/2}\left(\frac{1}{m_0}\right)(\mu-\hat{\mu}_{t-1})\right\|^2 \leqslant \beta_t\right\}.$$

Moreover, the contribution of the complement of the clean event to the overall regret is negligible. Plugging this choice of β_t into the bound of Lemma 2.5.20 completes the proof of Theorem 2.3.5.

Chapter 3

Beyond Performative Prediction

In Chapter 2 we developed a framework for reasoning about prediction when it affects and alters the phenomena it aims to describe. We introduced two solution concepts—performative stability and performative optimality—and studied algorithms for finding these equilibria.

In this chapter we take a step back and reevaluate the core of the performative prediction framework. In particular, we argue that there are ubiquitous scenarios where the act of prediction alters the patterns it aims to describe, but in a way that is different than—arguably even opposed to—the performative prediction formalism. These scenarios will be particularly common in the context of *online platforms*, whose distinctive feature is dominant computational power and abundant data resources.

To give an example of an interaction that is not adequately described by performative prediction, consider ride-sharing platforms that deploy algorithms for determining travel fare as a function of trip length and relevant traffic conditions. These pricing mechanisms are frequently updated based on the current supply and demand, and in particular a dip in the supply of drivers triggers so-called surge pricing. Möhlmann and Zalmanson [128] observed that drivers occasionally coordinate a massive deactivation of drivers from the platform, artificially lowering driver supply, only to get back on the platform after some time has passed and the prices have surged. In this example the platform's pricing algorithm *reacts* to the drivers' action. In other words, the algorithm responds to the population around which it operates. In contrast, a core feature of performative prediction is that the population—abstracted away via a distribution map—responds to the decision-making algorithm.

In reality the interaction between decision-making algorithms and the population they serve is two-way, and in this chapter we focus on the direction not captured in Chapter 2. Unlike in performative prediction, where the learner selects a predictive model described by a parameter vector θ and the population responds with a distribution $\mathcal{D}(\theta)$, we study settings where a population selects a data distribution $\mathcal{D}$ and the learner responds with a model $\theta(\mathcal{D})$. Like in the ride-sharing example, we focus on interactions driven by strategic incentives.

The material in this chapter is based on works co-authored with Moritz Hardt, Michael I. Jordan, Eric Mazumdar, Celestine Mendler-Dünner, and S. Shankar Sastry [75, 194].

3.1 The role of order of play

Individuals interacting with a decision-making algorithm often adapt strategically to the decision rule in order to achieve a desirable outcome. While such strategic adaptation might increase the individuals' utility, it also breaks the statistical patterns that justify the decision rule's deployment. This widespread phenomenon, often known as Goodhart's law, can be summarized as: "When a measure becomes a target, it ceases to be a good measure" [156].

A growing body of work known as *strategic classification* [21, 44, 76] models this phenomenon as a two-player game in which a decision-maker "leads" and strategic agents subsequently "follow." Specifically, the decision-maker first deploys a decision rule, and the agents then take a strategic action so as to optimize their outcome according to the deployed rule, subject to natural manipulation costs. For example, a bank might make lending decisions using applicants' credit scores. Knowing this mechanism, loan applicants might sign up for a large number of credit cards in an effort to strategically increase their credit score at little effort. As discussed in Chapter 2, strategic classification is a special case of performative prediction.

One of the main goals in strategic classification is to develop strategy-robust decision rules; that is, rules that remain meaningful even after the agents have adapted to them. Recent work has studied strategies for finding such rules through *repeated interactions* between the decision-maker and the agents [8, 38, 49]. In particular, the decision-maker sequentially deploys different rules, and for each they observe the population's response. Under regularity conditions, over time the decision-maker can find the optimal solution, defined as the rule that minimizes the decision-maker's loss *after* the agents have responded to the rule.

With the emergence of online platforms such as social media and e-commerce sites, repeated interactions between decision-makers and the population have become ever more prevalent. Online platforms continuously monitor user behavior and update pricing algorithms, recommendation systems, and popularity rankings accordingly. Users, on the other hand, take actions to ensure favorable outcomes in the face of these updates.

A distinctive feature of online platforms is the decision-maker's dominant computational power and abundant data resources, allowing the platform to react to any change in the agents' behavior virtually instantaneously. For example, if fake news content changes over time, automated algorithms can quickly detect this and retrain the classifier to incorporate the shift. It has been observed [see, e.g., 25, 40, 128] that, when faced with such "reactive" algorithms, strategic agents tend to take actions that *anticipate* the algorithm's response. That is, through repeated interactions, agents aim to find actions that maximize the agents' utility *after* the decision-maker has responded to these actions. This suggests that the order of play in strategic interactions can in fact be *reversed*, such that the agents "lead" while the decision-maker "follows."

We argue that the order of play in strategic classification is fundamentally tied to the relative *update frequencies* at which the decision-maker and the strategic agents adapt to each other's actions. We show that, by tuning their update frequency appropriately, the decision-maker can select the order of play in the underlying game. Furthermore, in natural settings we show that allowing the strategic agents to play first in the game can actually be preferable for *both* the decision-maker and the agents. This is contrary to the order of play previously studied in the literature, whereby the decision-maker is always assumed to make the first move.

3.1.1 Model

Throughout we denote by $z = (x, y)$ the feature–label pairs corresponding to the strategic agents' data. We assume that the decision-maker chooses a model parameterized by $\theta \in \Theta \subseteq \mathbb{R}^d$, where Θ is convex and closed, and that their loss is measured via a convex loss function $\ell(z; \theta)$. The strategic agents measure loss via a function $r(z; \theta)$ and, collectively, they form a distribution in the family $\{\mathcal{D}_\mu : \mu \in \mathcal{M} \subseteq \mathbb{R}^m\}$, where $\mathcal{M}$ is convex and closed. Here, μ denotes the aggregate summary of all agents' actions. The data observed by the decision-maker is $\mathcal{D}_\mu$ and as such varies depending on the agents' aggregate action μ.

We denote $L(\mu, \theta) = \mathbb{E}_{z\sim\mathcal{D}_\mu} \ell(z; \theta)$, and $R(\mu, \theta) = \mathbb{E}_{z\sim\mathcal{D}_\mu} r(z; \theta)$. With this, the agents' best response is given by $\mu_{\mathrm{BR}}(\theta) = \arg\min_\mu R(\mu, \theta)$ and the decision-maker's best response is given by $\theta_{\mathrm{BR}}(\mu) = \arg\min_\theta L(\mu, \theta)$. We assume that the best responses for both players are always unique.

If the decision-maker acts as the leader in the game, their incurred *Stackelberg risk* is equal to $\mathrm{SR}_L(\theta) = L(\mu_{\mathrm{BR}}(\theta), \theta)$. Similarly, we let $\mathrm{SR}_R(\mu) = R(\mu, \theta_{\mathrm{BR}}(\mu))$ denote the Stackelberg risk of the agents when they lead in the game. We let θ_{SE} and μ_{SE} denote the decision-maker's and strategic agents' equilibrium, respectively: $\theta_{\mathrm{SE}} = \arg\min_\theta \mathrm{SR}_L(\theta)$ and $\mu_{\mathrm{SE}} = \arg\min_\mu \mathrm{SR}_R(\mu)$. We assume that each equilibrium is unique. Note that the two players cannot compute their respective equilibrium “offline”, as we do not assume they have access to the other player's loss function.

As discussed earlier, we assume that there is an underlying timescale according to which the agents re-evaluate their features. Specifically, after each time interval of fixed length, the agents observe the currently deployed model, as well as their loss according to that model, and possibly modify their features accordingly. The decision-maker, aware of the agents' timescale, can choose to be *proactive*, meaning they choose an update frequency slower than that of the agents, or *reactive*, meaning they choose a higher update frequency. This power asymmetry that allows the decision-maker to choose a timescale is characteristic of online platforms with abundant resources.

We use the term *epoch* to refer to a period between two updates of the *slower* player (which player is the slower one is up to the decision-maker). In particular, the t-th epoch starts with a single update of the slower player, followed by $\tau \in \mathbb{N}$ updates of the faster player. The rate τ is fixed.

We use θ_t and μ_t to denote the iterate of the decision-maker and the strategic agents, respectively, *at the end* of epoch t. Furthermore, for the faster player, we use double-indexing to denote the within-epoch iterates. For example, if the decision-maker is the faster player, we use $\{\theta_{t,j}\}_{j=1}^\tau$ to denote their iterates within epoch t. Note that $\theta_{t,\tau} \equiv \theta_t$. We also let $\bar{\theta}_t = \frac{1}{\tau}\sum_{j=1}^\tau \theta_{t,j}$. We adopt similar notation when the agents have a higher update frequency.

Rational agents in the face of varying update frequencies

Adopting the distinction between reactive and proactive decision-makers, it is crucial to re-evaluate what it means for the strategic agents to behave rationally. We argue that rational behavior must depend on the relative update frequencies of the decision-maker and the agents.

As a running toy example, consider a decision-maker building a model with the goal of distinguishing between spam and legitimate emails. The population of strategic agents aims to craft emails that bypass the decision-maker's spam filter. Here, μ could determine the number of words in an email, types of words used, etc. The loss $R(\mu, \theta)$ could be some decreasing function of the number of daily clicks on email content, given spam filter θ and emails crafted according to μ. In the following discussion assume that the timescales of the decision-maker and the agents have a significant separation: the decision-maker is either "significantly faster" or "significantly slower." As we will make more formal later on, our results will generally assume a sufficiently large separation between the timescales. In the following paragraphs we informally describe rational agent behavior in the context of update frequencies.

Proactive decision-maker. First, assume that the decision-maker is proactive, and suppose they deploy model θ. By definition, this model remains in place for a relatively long time, as observed by the agents. Then, by choosing features μ, the agents experience loss $R(\mu, \theta)$ during that period, and as a result the most rational decision is to choose features $\mu_{\mathrm{BR}}(\theta)$. In the running example, if θ is a spam filter that is in place for many months, it is rational for spammers to craft emails that are most likely to bypass filter θ. This is just the usual best response—as we alluded to earlier, when the decision-maker is proactive, our setup is similar to that of strategic classification.

Reactive decision-maker. Now assume that the decision-maker is reactive, and suppose the agents observe θ as the current model. Then, by setting μ, the agents do *not* experience loss $R(\mu, \theta)$. Rather, their loss is $R(\mu, \theta_{\mathrm{R}}(\mu))$, where $\theta_{\mathrm{R}}(\mu)$ denotes the decision-maker's *reaction* to the agents' choice μ. In the spam example, suppose that the decision-maker can aggregate and process data quickly, and retrains the spam filter every couple of hours. Moreover, suppose that the spammers adapt their emails only once per week. Then, the agents' loss after choosing μ (evaluated weekly) is determined by the number of clicks allowed by the updated filter $\theta_{\mathrm{R}}(\mu)$, *not* the old filter θ. Therefore, if the agents could predict $\theta_{\mathrm{R}}(\mu)$, the agents' optimal decision would be to choose $\arg\min_\mu R(\mu, \theta_{\mathrm{R}}(\mu))$. In other words, rather than choose the best response to θ, rational agents interacting with a reactive decision-maker would choose μ so that it triggers the *best possible reaction* from θ.

We formalize this intuitive behavior by assuming that the agents are *no-regret* learners [147]. This essentially means that their average regret vanishes as the number of actions grows. More formally, we assume the following behavior depending on the relative update frequencies:

- If the decision-maker is proactive, then for any θ_t, the agents' strategy ensures:

$$\frac{1}{\tau}\sum_{j=1}^{\tau}\mathbb{E}\,R(\mu_{t,j},\theta_t) - \min_{\mu} R(\mu,\theta_t) \to 0 \text{ as } \tau \to \infty. \tag{A6}$$

- If the decision-maker is reactive, then for any response function $\theta_{\mathrm{R}}(\mu)$, the agents' strategy ensures:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\,R(\mu_t,\theta_{\mathrm{R}}(\mu_t)) - \min_{\mu} R(\mu,\theta_{\mathrm{R}}(\mu)) \to 0 \text{ as } T \to \infty, \tag{A7}$$

whenever such a strategy exists. If the agents' loss is convex, the first condition can be satisfied by simple gradient descent. In fact, gradient descent would typically imply an even stronger guarantee, namely the convergence of the iterates, $\mu_{t,\tau} \to \mu_{\mathrm{BR}}(\theta_t)$. The second condition can be satisfied by various bandit strategies if $R(\mu,\theta_{\mathrm{R}}(\mu))$ is Lipschitz and $\mathcal{M}$ is bounded (and we will impose these conditions explicitly in the following section). That said, it seems hardly suitable to assume that the agents run a well-specified optimization procedure. For this reason, we will for the most part avoid making explicit algorithmic assumptions on the agents' strategy and our main takeaways will only rely on rational agent behavior in the limit, as in equations (A6) and (A7).

3.1.2 Learning dynamics

We study the limiting behavior of the interaction between the decision-maker and the strategic agents. We show that, by running classical optimization algorithms, the decision-maker can drive the interaction to a Stackelberg equilibrium with either player acting as the leader.

Convergence to decision-maker's equilibrium

In general, we do not expect the decision-maker to be able to compute derivatives of the function SR_L. For this reason, to achieve convergence to the decision-maker's equilibrium, we consider running a derivative-free method. One such solution is the "gradient descent without a gradient" algorithm of Flaxman et al. [62]. Past work [49] also considers this algorithm with the goal of optimizing SR_L, but it assumes instantaneous agent responses. In other words, it assumes query access to SR_L directly, while we consider perturbations due to imperfect agent responses.

Specifically, we let the decision-maker run the following update:

$$\phi_{t+1} = \Pi_\Theta(\phi_t - \eta_t \frac{d}{\delta} L(\bar{\mu}_t, \phi_t + \delta u_t) u_t), \text{ where } u_t \sim \mathrm{Unif}\left(\mathcal{S}^{d-1}\right). \tag{3.1}$$

Here, Π_Θ denotes the Euclidean projection, $\text{Unif}\left(\mathcal{S}^{d-1}\right)$ denotes the uniform distribution on the unit sphere in $\mathbb{R}^d$, η_t is a non-increasing step size sequence, and $\delta > 0$ is a fixed hyperparameter. The deployed model in the t-th epoch is set as $\theta_t = \phi_t + \delta u_t$.[1]

We provide convergence guarantees assuming that the decision-maker's Stackelberg risk SR_L is convex. While this condition doesn't follow from convexity of the loss $\ell(z;\theta)$ alone, previous work has established conditions for convexity of this objective for different learning problems and agent utilities [49, 125]. For example, in the linear and logistic regression examples discussed in the following section, the decision-maker's Stackelberg risk will be convex.

Theorem 3.1.1. *Denote by D_Θ the diameter of Θ, and suppose that $|L(\mu,\theta)| \leqslant B$ for all μ, θ. Furthermore, suppose that SR_L is convex and β-Lipschitz and $L(\mu,\theta)$ is β_μ-Lipschitz in the first entry for all θ. Then, if the decision-maker runs update* (3.1) *with $\eta_t = \eta_0 d^{-\frac{1}{2}} t^{-\frac{3}{4}}$ and $\delta = \delta_0 d^{\frac{1}{2}} T^{-1/4}$, it holds that*

$$\sum_{t=1}^{T} (\mathbb{E}[\text{SR}_L(\theta_t)] - \text{SR}_L(\theta_{\text{SE}})) \leqslant \left(\frac{D_\Theta^2}{2\eta_0} + \frac{2B^2}{\delta_0^2}\right)\sqrt{d}T^{3/4} + \beta_\mu D_\Theta \sum_{t=1}^{T} \mathbb{E}\|\bar{\mu}_t - \mu_{\text{BR}}(\theta_t)\|_2.$$

Moreover, assuming that the agents are rational (A6) *and $\mathcal{M}$ is compact, we have*

$$\lim_{\tau\to\infty} \sum_{t=1}^{T} (\mathbb{E}[\text{SR}_L(\theta_t)] - \text{SR}_L(\theta_{\text{SE}})) \leqslant \left(\frac{D_\Theta^2}{2\eta_0} + \frac{2B^2}{\delta_0^2}\right)\sqrt{d}T^{3/4}. \tag{3.2}$$

Remark 3.1.1. *For Theorem 3.1.1, we assume that the agents are rational in a relatively weak sense, by assuming no-regret behavior. Often, however, we expect the agents' strategy to achieve* iterate convergence*, and not just vanishing regret. More precisely, it makes sense to expect $\mu_{t,\tau} \to \mu_{\text{BR}}(\theta_t)$ as $\tau \to \infty$. For example, this guarantee is achieved by gradient descent in a variety of settings. In that case, the decision-maker can simply use the* last *iterate instead of the average one:*

$$\phi_{t+1} = \Pi_\Theta(\phi_t - \eta_t \frac{d}{\delta} L(\mu_t, \phi_t + \delta u_t) u_t), \text{ where } u_t \sim \text{Unif}\left(\mathcal{S}^{d-1}\right). \tag{3.3}$$

Similarly, $\mathbb{E}\|\bar{\mu}_t - \mu_{\text{BR}}(\theta_t)\|_2$ would be replaced by $\mathbb{E}\|\mu_t - \mu_{\text{BR}}(\theta_t)\|_2$ in the bound of Theorem 3.1.1.

In some cases, the additional regret due to imperfect agent responses does not alter the asymptotic rate at which the decision-maker accumulates regret even if the epoch length τ is constant and does not grow with T. To illustrate this point, we consider strategic agents that follow the gradient-descent direction on a possibly nonconvex objective with enough

[1] Technically, this assumes that we can deploy a model in a δ-ball around Θ. Another solution would be to use a projection onto a small contraction of Θ in equation (3.1). This is a minor technical hurdle common in the literature. The rate in Theorem 3.1.1 is unaffected by the choice of solution to this technical point.

curvature. More precisely, we assume that for all θ, $R(\mu, \theta)$ satisfies the Polyak-Łojasiewicz (PL) condition:

$$\gamma(R(\mu,\theta) - \min_{\mu\in\mathcal{M}} R(\mu,\theta)) \leqslant \frac{1}{2}\|\nabla_\mu R(\mu,\theta)\|_2^2,$$

for some parameter $\gamma > 0$. Suppose that the agents' update is computed as:

$$\mu_{t,j+1} = \mu_{t,j} - \eta_\mu \nabla_\mu R(\mu_{t,j}, \theta_t), \tag{3.4}$$

where $\eta_\mu > 0$ is a constant step size and $\mu_{t,0} = \mu_{t-1,\tau}$. In this case, gradient descent achieves last-iterate convergence and hence we assume that the decision-maker uses the update in equation (3.3).

Theorem 3.1.2. *Assume the conditions of Theorem 3.1.1. In addition, suppose that $R(\mu,\theta)$ is β_μ^R-smooth in μ for all θ and satisfies the PL condition with parameter γ, and $\mu_{\mathrm{BR}}(\theta)$ is β_{BR}-Lipschitz in θ. Assume that the strategic agents run update* (3.4) *with $\eta_\mu < \frac{1}{\beta_\mu^R}$. Further, suppose the epoch length is chosen so that $\tau > \log(\beta_\mu^R/\gamma)/\log(1/(1-\gamma\eta_\mu))$. Then, for some constant $\alpha(\tau) \in (0,1)$, we have*

$$\sum_{t=1}^{T} \mathbb{E}\,\|\mu_t - \mu_{\mathrm{BR}}(\theta_t)\|_2 \leqslant \frac{\|\mu_0 - \mu_{\mathrm{BR}}(\theta_0)\|_2 + \frac{4\beta_{\mathrm{BR}} B \eta_0}{\delta_0}\sqrt{T}}{1-\alpha(\tau)}.$$

Therefore, the decision-maker's regret is $O(\sqrt{d}T^{3/4})$ even with a constant epoch length. This result crucially depends on the fact that the optimization problems that the agents solve in neighboring epochs are coupled through $\mu_{t,0} = \mu_{t-1,\tau}$. If $\mu_{t,0}$ were reinitialized arbitrarily in each epoch, the extra regret would be linear in T given constant epoch length.

Convergence to strategic agents' equilibrium

Now we analyze the case when the decision-maker is reactive. Given a large enough gap in update frequencies—that is, a large enough epoch length τ—the decision-maker can converge to their best response to the current iterate μ_t between any two actions of the agents. The most natural choice for achieving this is to run standard gradient descent, $\theta_{t,k+1} = \theta_{t,k} - \eta_k \nabla_\theta L(\mu_t, \theta_{t,k})$. In what follows we provide asymptotic guarantees assuming that the decision-maker runs any algorithm that achieves iterate convergence. This condition can be satisfied by gradient descent in a variety of settings. Formally, we assume that for any fixed μ_t, the decision-maker's strategy ensures

$$\|\theta_{t,\tau} - \theta_{\mathrm{BR}}(\mu_t)\|_2 \to_p 0, \tag{3.5}$$

as $\tau \to \infty$. Here, $\to_p$ denotes convergence in probability.

We first observe that, in the limit as τ grows, the agents' accumulated risk is equal to their accumulated *Stackelberg risk* at all the actions played so far. This simply follows by continuity.

Lemma 3.1.1. *Suppose that the decision-maker achieves iterate convergence* (3.5) *and R is continuous in the second argument. Then, for all $T \in \mathbb{N}$, $\lim_{\tau\to\infty}\sum_{t=1}^{T}\mathbb{E}\,R(\mu_t,\theta_t) = \sum_{t=1}^{T}\mathbb{E}\,\mathrm{SR}_R(\mu_t)$.*

In other words, in every epoch the agents essentially play a Stackelberg game in which they lead and the decision-maker follows. This holds regardless of whether the agents behave rationally. If they do behave rationally (condition (A7)), we show that both the agents' and the decision-maker's average regret with respect to $(\mu_{\mathrm{SE}}, \theta_{\mathrm{BR}}(\mu_{\mathrm{SE}}))$ vanishes if the agents' updates are continuous. To formalize this, suppose that for all $t \in \mathbb{N}$, the agents set $\mu_{t+1} = D_{t+1}(\mu_1, \theta_1, \dots, \mu_t, \theta_t, \xi_{t+1})$, where D_{t+1} is some fixed map and ξ_{t+1} is a random variable independent of $\{(\mu_i, \theta_i)\}_{i\leqslant t}$. We include ξ_{t+1} as an input to allow randomized strategies. Then, we will say that the agents' updates are *continuous* if D_{t+1} is continuous in the first $2t$ coordinates for all $t \in \mathbb{N}$.

Theorem 3.1.3. *Suppose that the agents' updates are continuous and rational* (A7)*, and that $\mathcal{M}$ is compact. Further, suppose that the decision-maker achieves iterate convergence* (3.5) *and* SR_R *and* SR_L *are Lipschitz. Then, it holds that*

$$\lim_{T\to\infty}\lim_{\tau\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\,\mathrm{SR}_R(\mu_t) - \mathrm{SR}_R(\mu_{\mathrm{SE}}) = 0, \quad \lim_{T\to\infty}\lim_{\tau\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\,L(\mu_t,\theta_t) - L(\mu_{\mathrm{SE}}, \theta_{\mathrm{BR}}(\mu_{\mathrm{SE}})) = 0.$$

3.1.3 Preferred order of play

While we have shown that the decision-maker can tune their update frequency to achieve either order of play in the Stackelberg game, it remains to understand which order of play is preferable for the decision-maker and the strategic agents. In the following examples, we illustrate that in classic learning settings both players can prefer the order when the *agents lead*. This suggests that the natural and overall more desirable order of play is sometimes reversed compared to the order usually studied.

At first, it might seem counterintuitive that the decision-maker could prefer to follow. To get some intuition for why following might be preferred to leading, recall that in zero-sum games *following is never worse*. In particular, suppose $R(\mu,\theta) = -L(\mu,\theta)$. Then, the basic min-max inequality says

$$L(\mu_{\mathrm{SE}}, \theta_{\mathrm{BR}}(\mu_{\mathrm{SE}})) = \max_{\mu}\min_{\theta} L(\mu,\theta) \leqslant \min_{\theta}\max_{\mu} L(\mu,\theta) = L(\mu_{\mathrm{BR}}(\theta_{\mathrm{SE}}), \theta_{\mathrm{SE}}),$$

with equality if and only if a *Nash* equilibrium exists. Therefore, if a Nash equilibrium does not exist, following is strictly preferred.

Since strategic classification is typically not a zero-sum game, we look at two common learning problems and analyze the preferred order of play.

Linear regression

Suppose that the agents' non-strategic data, (x_0, y), where x_0 is a feature vector and y the outcome, is generated according to

$$x_0 \sim \mathcal{D}_0,\ y = x_0^\top \beta + \xi,$$

where $\mathcal{D}_0$ is a zero-mean distribution such that $\mathbb{E}_{x_0 \sim \mathcal{D}_0} x_0 x_0^\top = I$, $\beta \in \mathbb{R}^d$ is an arbitrary fixed vector, and ξ has mean zero and finite variance σ^2. We denote the joint distribution of (x_0, y) by $\mathcal{D}_0$.

Recall that we use z to denote the pair (x, y). Suppose that the decision-maker runs standard linear regression with the squared loss:

$$\ell(z;\theta) = \frac{1}{2}(y - x^\top \theta)^2.$$

The agents aim to maximize their predicted outcome, $r(z;\theta) = -\theta^\top x$, subject to a fixed budget on feature manipulation—they can move to any x at distance at most B from their original features x_0: $\|x - x_0\|_2 \leqslant B$. A similar model is considered by Kleinberg and Raghavan [99] and Chen et al. [38]. More precisely, we let $\mathcal{M} = \{\mu \in \mathbb{R}^d : \|\mu\|_2 \leqslant B\}$ and define $\mathcal{D}_\mu$ to be the distribution of (x, y), where $(x_0, y) \sim \mathcal{D}_0$ and $x = x_0 + \mu$. Then, $R(\mu, \theta) = \mathbb{E}_{z \sim \mathcal{D}_\mu} r(z;\theta) = -\mu^\top \theta$ and $L(\mu, \theta) = \mathbb{E}_{z \sim \mathcal{D}_\mu} \ell(z;\theta)$.

We prove that both the decision-maker and the agents prefer the agents' equilibrium.

Proposition 3.1.1. *Assume the linear regression setup described above. Then, we have*

$$\frac{\sigma^2}{2} + \frac{\|\beta\|_2^2 \min(1,B)^2}{2(1+\min(1,B)^2)} = L(\mu_{\mathrm{SE}}, \theta_{\mathrm{BR}}(\mu_{\mathrm{SE}})) \leqslant \mathrm{SR}_L(\theta_{\mathrm{SE}}) = \frac{\sigma^2}{2} + \frac{\|\beta\|_2^2 B^2}{2(1+B^2)},$$
$$-\frac{\|\beta\|_2 \min(1,B)}{1+\min(1,B)^2} = \mathrm{SR}_R(\mu_{\mathrm{SE}}) \leqslant R(\mu_{\mathrm{BR}}(\theta_{\mathrm{SE}}), \theta_{\mathrm{SE}}) = -\frac{\|\beta\|_2 B}{1+B^2}.$$

When $B \leqslant 1$, the losses implied by the two scenarios are the same, while when $B > 1$, having the agents lead is strictly better for both players. Moreover, the strategic agents' manipulation cost is no higher when they lead: $\|\mu_{\mathrm{SE}}\|_2 \leqslant \|\mu_{\mathrm{BR}}(\theta_{\mathrm{SE}})\|_2$.

Logistic regression

Next we consider a classification example. Suppose that the non-strategic data (x_0, y) is sampled according to a base joint distribution $\mathcal{D}_0$ supported on $\mathbb{R}^d \times \{0, 1\}$. Unlike in the linear regression example, we place no further constraint on $\mathcal{D}_0$.

We assume that the decision-maker trains a logistic regression classifier:

$$\ell(z;\theta) = -yx^\top \theta + \log(1 + e^{x^\top \theta}).$$

The agents with $y = 0$ can manipulate their features to increase the probability of being positively labeled. A similar setup is considered by Dong et al. [49]. As in the previous

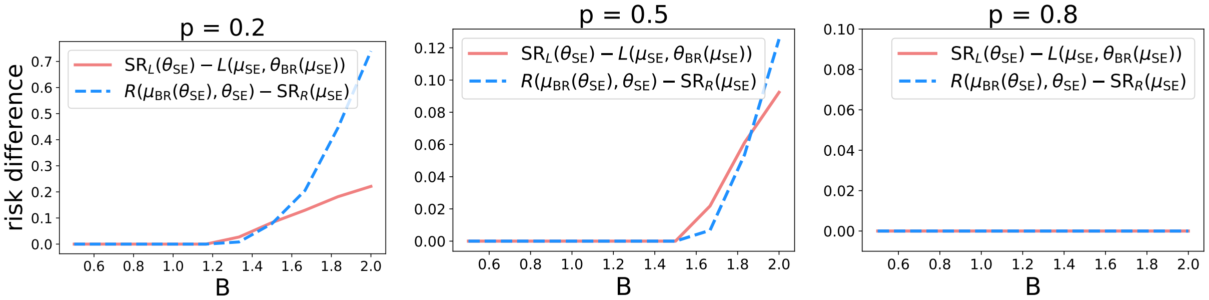


Figure 3.1: Difference in decision-maker's and agents' risk implied by the two Stackelberg equilibria, for different values of B and p.

example, the agents have a limited budget to change their features: if their non-strategic features are x_0, they can move to any x which is at distance at most B from x_0, $\|x - x_0\|_2 \leqslant B$. Thus, we set $\mathcal{M} = \{\mu \in \mathbb{R}^d : \|\mu\|_2 \leqslant B\}$ and denote by $\mathcal{D}_\mu$ the joint distribution of (x, y) where $(x_0, y) \sim \mathcal{D}_0$ and $x = x_0 + \mu \mathbf{1}\{y = 0\}$. We let $R(\mu, \theta) = -\mu^\top \theta$ and $L(\mu, \theta) = \mathbb{E}_{z \sim \mathcal{D}_\mu} \ell(z; \theta)$.

Proposition 3.1.2. *Assume the logistic regression setup described above. Then, we have*

$$L(\mu_{\mathrm{SE}}, \theta_{\mathrm{BR}}(\mu_{\mathrm{SE}})) \leqslant \mathrm{SR}_L(\theta_{\mathrm{SE}}) \text{ and } \mathrm{SR}_R(\mu_{\mathrm{SE}}) \leqslant R(\mu_{\mathrm{BR}}(\theta_{\mathrm{SE}}), \theta_{\mathrm{SE}}).$$

There exist configurations of parameters such that the inequalities in Proposition 3.1.2 are strict, meaning that both players strictly prefer the agents to lead. We illustrate this empirically. In Figure 3.1 we generate non-strategic data according to $y \sim \mathrm{Bern}(p)$ and $x_0|y \sim N(4y - 2, 1)$ and plot the difference in risk between the two equilibria for the decision-maker and the agents, for varying B and p. For large p and small B, we see no difference between the equilibria. However, as p decreases and B increases, it becomes suboptimal for both players if the decision-maker leads.

3.1.4 Experiments

As proof of concept, we demonstrate our theoretical findings empirically in a simulated logistic regression setting. The non-strategic data is generated as

$$y \sim \mathrm{Bern}(p) \text{ and } x_0|y \sim \mathcal{N}((2y - 1)\alpha, I).$$

In other words, $x_0|y = 1 \sim \mathcal{N}(\alpha, I)$ and $x_0|y = 0 \sim \mathcal{N}(-\alpha, I)$.

In the first set of experiments, we adopt the model from Section 3.1.3 where agents are constrained in how they modify their features. In the second set of experiments we adopt a model more akin to that of Dong et al. [49] where the negatively classified agents are penalized from deviating from their true features.

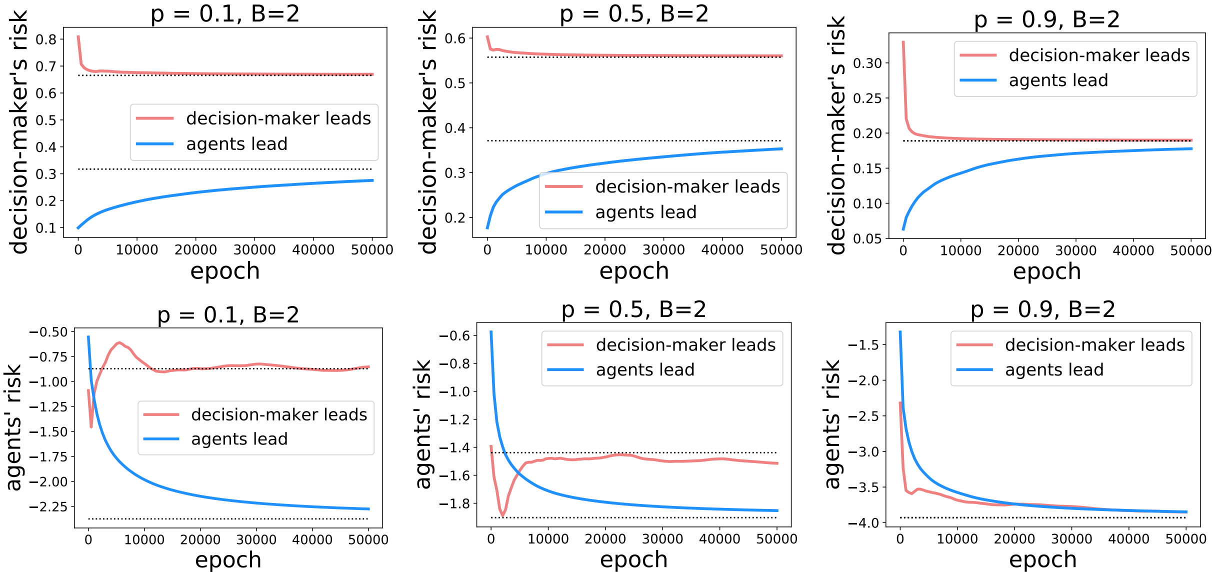


Figure 3.2: Decision-maker's and agents' average running risk for varying p and $B = 2$. The dotted lines denote the loss at the respective equilibria. For $p = 0.9$, the decision-maker's equilibrium and the agents' equilibrium coincide and the curves converge to the same value.

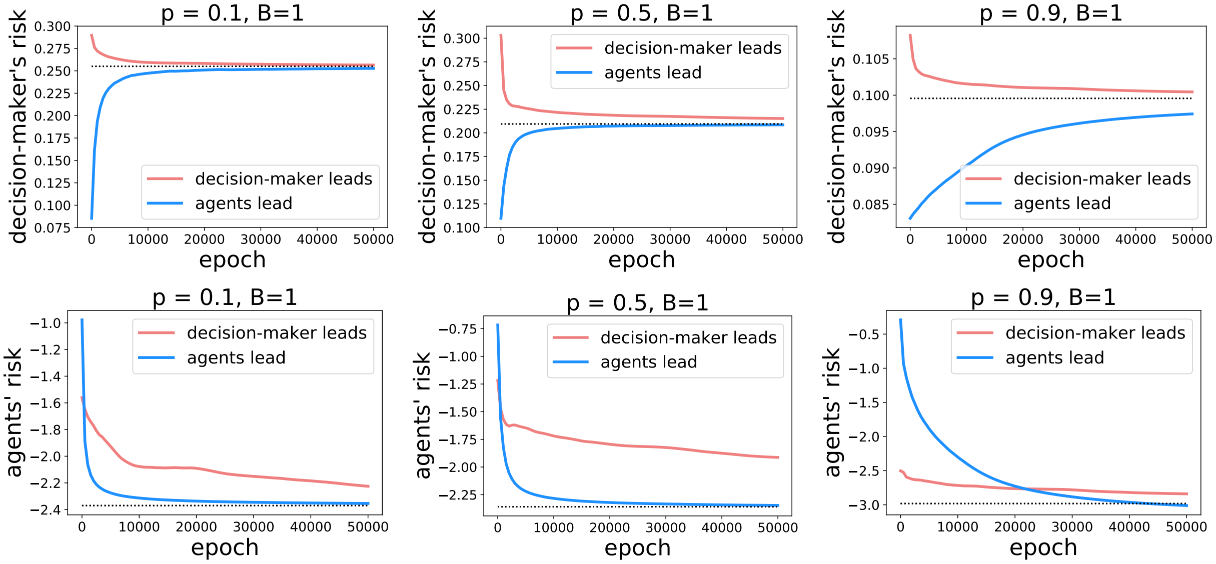


Figure 3.3: Decision-maker's and agents' average running risk for varying p and $B = 1$. The dotted lines denote the loss at the respective equilibria. The decision-maker's equilibrium and the agents' equilibrium coincide everywhere and the curves converge to the same value.

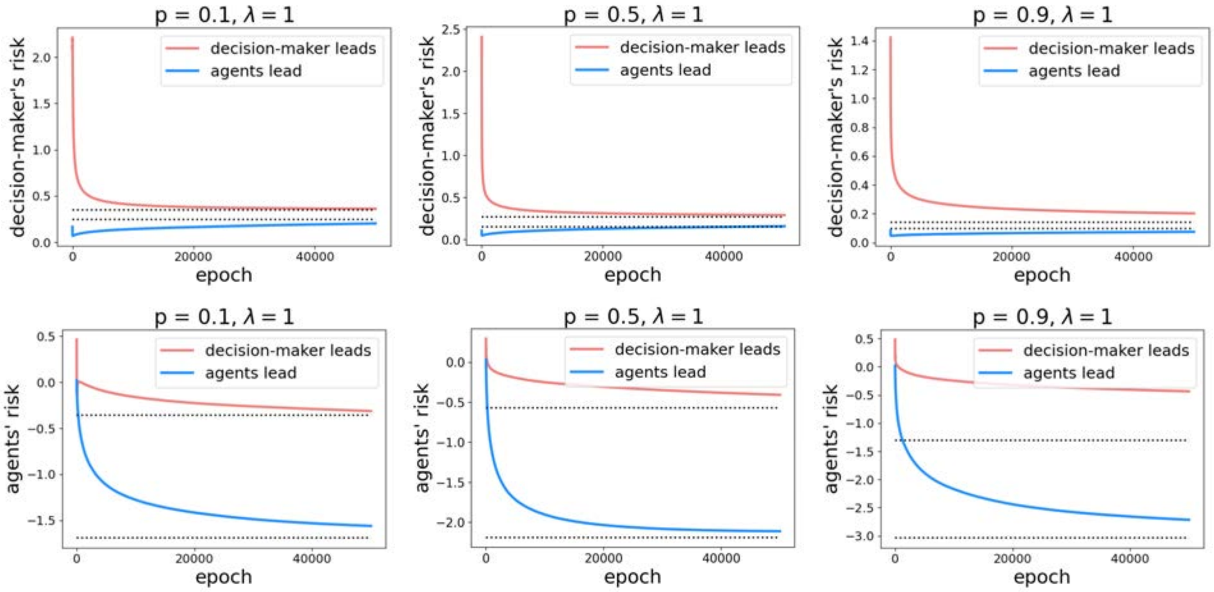


Figure 3.4: Decision-maker's and agents' average running risk for varying p and $\lambda = 1$. The dotted lines denote the loss at the respective equilibria.

In both settings, first we let the decision-maker lead and the agents follow, and then we switch the roles. For both orders of play, the slower player runs the derivative-free update (3.3), and the faster player runs standard (projected) gradient descent. To be able to analyze the long-run behavior, we also numerically approximate the Stackelberg risks of the decision-maker and the strategic agents and find the global minima which correspond to the decision-maker's and agents' equilibria respectively.

Agents with constraints

To begin, we verify our theoretical findings from Section 3.1.3. We generate 100 samples, fix $\alpha = 2$, $d = 1$, and vary B and p. We run the interaction for a total of $T = 50000$ epochs, with each epoch of length $\tau = 200$.

In Figure 3.2 and Figure 3.3 we plot the decision-maker's and the agents' average running risk against the number of epochs, for the two different orders of play, for $B = 2$ and $B = 1$, respectively. For $p \in \{0.1, 0.5\}$ and $B = 2$, we observe a clear gap between leading and following, the agents leading being the preferred order for both players. For $p = 0.9$ or $B = 1$, the two equilibria coincide asymptotically; however, generally we find that the two players still prefer the agents to lead even after a finite number of epochs.

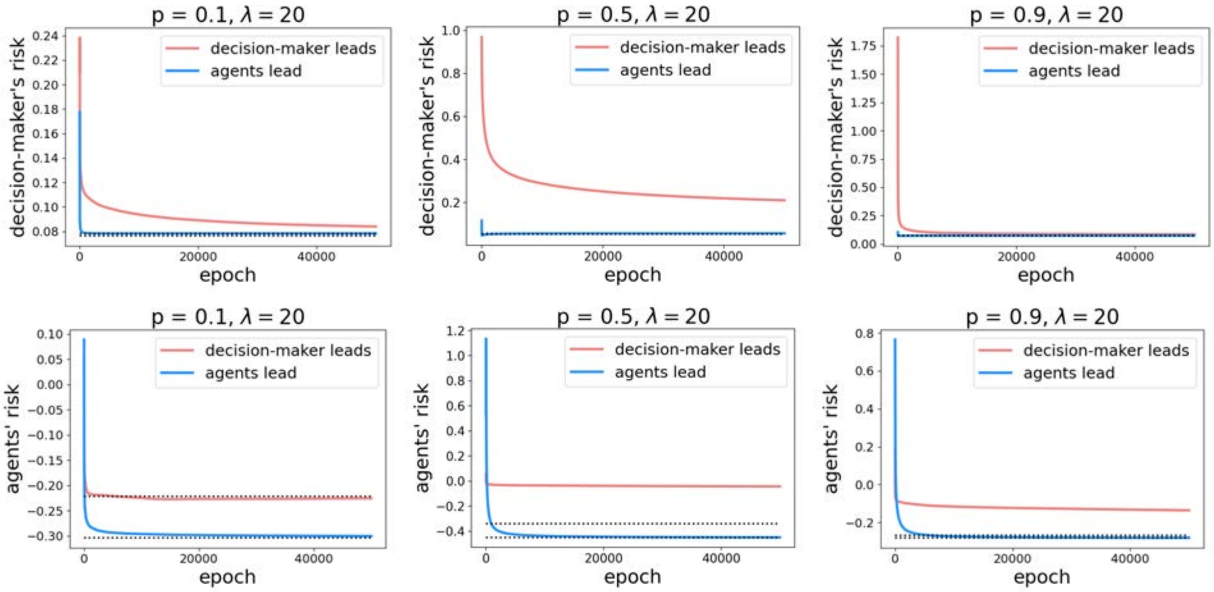


Figure 3.5: Decision-maker's and agents' average running risk for varying p and $\lambda = 20$. The dotted lines denote the loss at the respective equilibria. For $p = 0.9$, the decision-maker's equilibrium and the agents' equilibrium coincide and the curves converge to the same value.

Agents with costly deviations

In this section, we verify our findings on a model where the decision-maker's problem is the same logistic regression problem posed in Section 3.1.3, but the strategic agents are penalized for deviating from their true features. In particular, the agents' risk R takes the form: $R(\mu, \theta) = \frac{\lambda}{2}\|\mu\|^2 - \mu^T\theta$. We note that although this setup is conceptually very similar to that in Section 3.1.3 (increasing λ can be seen as shrinking the constraint set), it allows us to highlight that the experimental results are not caused by interactions with the constraints. Further, this setup is more readily comparable to previous models studied in, e.g., [49].

We generate 100 samples in $\mathbb{R}^2$, fix $\alpha = 1.5[1, 1]^\top$, and vary λ and p. We run the interaction for a total of $T = 50000$ epochs, with each epoch of length $\tau = 100$. In Figure 3.4 and Figure 3.5 we plot the decision-maker's and the agents' average running risk against the number of epochs, for the two different orders of play and for $\lambda = 1$ and $\lambda = 20$ respectively.

In Figure 3.4 we observe a gap between the decision-maker's risk at their Stackelberg equilibrium and at the agents', and see that the decision-maker consistently prefers the agents leading. The agents consistently prefer leading as well, meaning that *both* sides prefer if the order of play is flipped. In Figure 3.5 we observe that as λ and p increase, the gap between the two equilibria shrinks and disappears entirely when $p = 0.9$ and $\lambda = 20$. This is similar to the behavior seen in the constrained agent problem where shrinking the constraint set gives rise to Nash equilibria where neither player strictly prefers leading or following.

3.2 Algorithmic collective action

Throughout the gig economy, numerous digital platforms algorithmically profile, control, and discipline workers that offer on-demand services to consumers. Data collection and predictive modeling are critical for a typical platform's business as machine learning algorithms power ranking, scoring, and classification tasks of various kinds [70, 150, 184].

Troves of academic scholarship document the emergence and preponderance of precarity in the gig economy. [181] argue that platform-based algorithmic control can lead to "low pay, social isolation, working unsocial and irregular hours, overwork, sleep deprivation and exhaustion." This is further exacerbated by "high levels of inter-worker competition with few labor protections and a global oversupply of labor relative to demand." In response, there have been numerous attempts by gig workers to organize in an effort to reconfigure working conditions. A growing repertoire of strategies, as vast as it is eclectic, uses both physical and digital means towards this goal. Indeed, workers have shown significant ingenuity in creating platform-specific infrastructure, such as their own mobile apps, to organize the labor side of the platform [34, 140]. Yet, "the upsurge of worker mobilization should not blind us to the difficulties of organizing such a diverse and spatially dispersed labor force." [167]

Beyond the gig economy, evidence of consumers seeking to influence the algorithms that power a platform's business is abundant. Examples include social media users attempting to suppress the algorithmic upvoting of harmful content by sharing screenshots rather than original posts [25], or individuals creating bots to influence crowd-sourced navigation systems [155]. The ubiquity of such strategic attempts calls for a principled study of how coordinated groups can wield control over the digital platforms to which they contribute data.

In this section, we study how a collective of individuals can algorithmically strategize against a learning platform. We envision a collective that pools the data of participating individuals and executes an algorithmic strategy by instructing participants how to modify their own data. The firm in turn solves a machine learning problem over the resulting data. The goal of the collective is to redirect the firm's optimization towards a solution that serves the collective. Notably, coordination is a crucial lever. When data are plentiful, a single individual lacks the leverage to unilaterally change the output of a learning algorithm; in contrast, we show that even small collectives can exert substantial influence.

3.2.1 Problem formulation

We study the strategic interaction of a firm operating a predictive system with a population of individuals. We assume that the the firm deploys a learning algorithm θ that operates on data points in a universe $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Each individual corresponds to a single data point $z \in \mathcal{Z}$, typically a feature–label pair. We model the population of individual participants as a distribution $\mathcal{D}_0$ over $\mathcal{Z}$.

We say that a fraction $\alpha > 0$ of the individuals form a *collective* in order to strategically respond to the firm's learning behavior. The collective agrees on a potentially randomized strategy $h : \mathcal{Z} \to \mathcal{Z}$ from a space of available strategies $\mathcal{H}$. The possible strategies $\mathcal{H}$ capture feasible changes to the data. For example, content creators on a video streaming platform may be indifferent between videos that differ only in a hidden watermark not visible to human viewers. Freelancers may be indifferent between two resumes that differ only in inconsequential syntactic details.

The firm therefore observes a mixture distribution

$$\mathcal{D} = \alpha\mathcal{D}^* + (1-\alpha)\mathcal{D}_0,$$

where we use $\mathcal{D}^*$ to denote the distribution of $h(z), z \sim \mathcal{D}_0$.

The collective strives to choose a strategy h so as to maximize a measure of success over the solution $f = \theta(\mathcal{D})$ chosen by the firm. Here, f is a mapping from features to labels, $f : \mathcal{X} \to \mathcal{Y}$. Given a strategy, we use $S(\alpha)$ to denote the level of success achieved by a collective of size α. The central question we study is how the success $S(\alpha)$ grows as a function of collective size α, and how large α needs to be in order to achieve a target success level.

Definition 3.2.1 (Critical mass)**.** *The critical mass for a target success level S^* is defined as the smallest α for which there exists a strategy such that $S(\alpha) \geqslant S^*$.*

Note that, although motivated from the perspective of labor, our formal model can also serve as a basis for studying collective action on the consumer side of digital platforms. Before presenting our results we briefly discuss why we focus on collective strategies.

Why collective action? By engaging in collective action, individuals can exert influence on the learning algorithm that they could not achieve by acting selfishly. In large-population settings such as online platforms, an individual contributes an infinitesimal fraction of the data used by the learning algorithm. Thus, under reasonable manipulation constraints, individual behavior is largely powerless in systematically changing the deployed model. Instead, individuals are limited to simple adversarial attacks or individual strategies that do not have lasting effects on the learning outcome. By coordinating individuals, however, collectives can wield enough power to steer learning algorithms towards desired goals. In subsequent sections we show that collectives can often do so while only representing a small fraction of the training data.

3.2.2 Collective action in classification

We start with classification under the assumption that the firm chooses an approximately optimal classifier on the data distribution $\mathcal{D}$.

Definition 3.2.2 (ε-optimal classifier)**.** *A classifier $f\colon \mathcal{X} \to \mathcal{Y}$ is ε-optimal under the distribution $\mathcal{D}$ if there exists a $\mathcal{D}'$ with $\mathrm{TV}(\mathcal{D},\mathcal{D}') \leqslant \varepsilon$ such that*

$$f(x) = \arg\max_{y\in\mathcal{Y}} \mathcal{D}'(y|x)\,.$$

Note that a 0-optimal classifier is the Bayes optimal classifier with respect to the zero–one loss function.

Under the above assumption, we focus on two general goals for the collective: *planting a signal* and *erasing a signal*.

Planting a signal

Assume the collective wants the classifier to learn an association between an altered version of the features $g(x)$ and a chosen target class y^*. Formally, given a transformation $g : \mathcal{X} \to \mathcal{X}$, the collective wants to maximize the following measure of success:

$$S(\alpha) = \mathbb{P}_{x \sim \mathcal{D}_0} \{f(g(x)) = y^*\} .$$

We call this objective "planting a signal" and $\mathcal{X}^* = \{g(x) \colon x \in \mathcal{X}\}$ the signal set. For example, $g(x)$ could be instance x with an inconsequential trigger (such as a video with an imperceptible watermark or a resume with a unique formatting) and y^* could be a label indicating that the instance is of high quality (e.g., a high-quality video or a highly qualified individual). As another example, the collective may have an altruistic goal to help individuals in a vulnerable subpopulation $\mathcal{X}_0 \subseteq \mathcal{X}$ achieve a desired outcome y^*. In this case, $g(x)$ could be a mapping from x to a randomly chosen instance in $\mathcal{X}_0$.

We provide natural strategies for planting a signal and characterize their success as a function of α. The key parameter that we identify as driving success is the *uniqueness* of the signal.

Definition 3.2.3 (ξ-unique signal)**.** *We say that a signal is ξ-unique if it satisfies*

$$\mathcal{D}_0(\mathcal{X}^*) \leqslant \xi.$$

In addition, success naturally depends on how suboptimal y^* is on the signal set under the base distribution. To formalize this dependence, we define the suboptimality gap of y^*:

$$\Delta = \max_{x \in \mathcal{X}^*} \left(\max_{y \in \mathcal{Y}} \mathcal{D}_0(y|x) - \mathcal{D}_0(y^*|x) \right) .$$

We consider two possibilities for the space of available strategies $\mathcal{H}$. First, we assume that the individuals can modify both features and labels. We call the resulting strategies *feature–label* strategies. Modifying features by, say, planting a trigger often comes at virtually no cost. Changing the label, however, may be hard, costly, or even infeasible. This is why we also consider *feature-only* strategies; such strategies only allow changes to features.

Feature–label signal strategy. We define the feature–label signal strategy as

$$h(x, y) = (g(x), y^*) . \tag{3.6}$$

The result below quantifies the success of this strategy in terms of the collective size and the uniqueness of the signal.

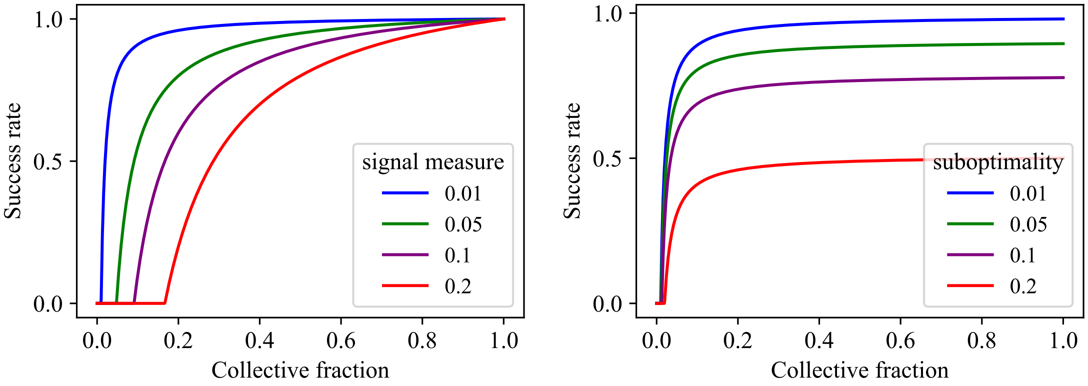


Figure 3.6: Illustration of the success rate predicted by Theorem 3.2.1. In the first we fix $\varepsilon = 0$ and vary ξ, and in the second we fix ξ and vary the classifier's suboptimality, ε. We upper bound Δ by one.

Theorem 3.2.1. *Consider the feature–label signal strategy and suppose that the signal is ξ-unique. Then, the success against an ε-optimal classifier is lower bounded by*

$$\mathrm{SR}(\alpha) \geqslant 1 - \frac{1-\alpha}{\alpha} \cdot \Delta \cdot \xi - \frac{\varepsilon}{1-\varepsilon}\,.$$

Rearranging the terms, we obtain an upper bound on the critical mass given a desired success probability (e.g., 90%).

Corollary 3.2.1. *Suppose the signal is ξ-unique. Then, the critical mass for achieving success* $\mathrm{SR}^* \in (0,1)$ *with feature–label strategies against an ε-optimal classifier is bounded by*

$$\alpha^* \leqslant \frac{\Delta \cdot \xi}{1 - \mathrm{SR}^* - \frac{\varepsilon}{1-\varepsilon} + \Delta \cdot \xi}. \tag{3.7}$$

Therefore, in order to achieve success it suffices to have a collective size proportional to the uniqueness of the signal and the suboptimality of y^* on the signal set, as long as these parameters are sufficiently small relative to the target error rate $1 - S^*$. This suggests that planting signals that are exceedingly “rare” under the base distribution can be done successfully by small collectives— a property of feature–label strategies that we empirically validate in Section 3.2.4.

In the next result we consider feature-only strategies. An impediment to the success of such strategies is the situation where two likelihoods $\mathcal{D}_0(x|y)$ and $\mathcal{D}_0(x|y')$ for distinct labels $y \neq y'$ have no overlapping support. In this case, there is no reason to expect that planting a signal in one class has any effect on the other class. This is the reason why we make one additional assumption that there exists a number $p > 0$ such that $\mathcal{D}_0(y^*|x) \geqslant p$ for all $x \in \mathcal{X}$.

Feature-only signal strategy. We define the feature-only signal strategy as

$$h(x,y)=\begin{cases}(g(x),y^*), & \text{if } y=y^*,\\ (x,y), & \text{otherwise.}\end{cases} \tag{3.8}$$

This strategy achieves a similar success rate as the feature–label strategy, but the success diminishes with the constant p.

Theorem 3.2.2. *Consider the feature-only signal strategy and suppose that the signal is ξ-unique. Further, suppose there exists $p>0$ such that $\mathcal{D}_0(y^*|x)\geqslant p, \forall x\in\mathcal{X}$. Then, the success against an ε-optimal classifier is lower bounded by*

$$\mathrm{SR}(\alpha)\geqslant 1-\frac{1-p}{p\alpha}\cdot\xi-\frac{\varepsilon}{1-\varepsilon}\,.$$

The critical mass for achieving a target success probability is thus bounded as follows.

Corollary 3.2.2. *Suppose the signal is ξ-unique. Then, the critical mass for achieving success* $\mathrm{SR}^*\in(0,1)$ *with feature-only strategies against an ε-optimal classifier is bounded by*

$$\alpha^*\leqslant\frac{1-p}{p}\frac{\xi}{1-\mathrm{SR}^*-\frac{\varepsilon}{1-\varepsilon}}. \tag{3.9}$$

Whenever the positivity constant p is smaller than 0.5, the critical mass (3.9) that guarantees success of feature-only strategies is at least as large as the critical mass (3.7) for feature–label strategies, as expected.

The positivity constant $p>0$ may be excessively small over the entire data universe. A standard fix to this problem is to restrict $\mathcal{D}_0$ to a subset where the constant is larger, and pay a penalty for the amount of truncation in the bound. For example, if there exists $R\subseteq\mathcal{X}$ such that $\mathcal{D}_0(R)\geqslant 99\%$, but the positivity constant over R is much larger than p, then one can obtain a more powerful version of Theorem 3.2.2.

Erasing a signal

Next, we assume the collective wants the classifier to be invariant under a transformation $g:\mathcal{X}\to\mathcal{X}$ of the features. In particular, the success is measured with respect to:

$$\mathrm{SR}(\alpha)=\mathbb{P}_{x\sim\mathcal{D}_0}\{f(x)=f(g(x))\}.$$

In other words, the collective wants the classifier to output the same predictions for all x and x' that have $g(x)=g(x')$. The map g can be thought of as a summary of x that removes some feature information. We call this objective "erasing a signal." For example, if the collective wants the deployed model to be insensitive to the value of a particular feature j^*, then it can use $g(x)=x'$ where $x'_j=x_j$ for $j\neq j^*$ and $x'_{j^*}=0$. The feature j^* could be the length of a video that content creators do not want to affect the ranking of the content, or it could be a sensitive demographic feature that a collective wants to be independent of the predicted label.

Erasure strategy. We define the erasure strategy as

$$h(x,y) = \left(x, \arg\max_{y\in\mathcal{Y}} \mathcal{D}_0(y|g(x))\right).$$

As before, the success of the strategy depends on problem-dependent quantities. In this case, the quantity of interest is the sensitivity of the labels to the erased signal. We capture this sensitivity in the parameter τ, defined as

$$\tau = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_0} \max_{y\in\mathcal{Y}} \left|\mathcal{D}_0(y|x) - \mathcal{D}_0(y|g(x))\right|.$$

Intuitively, τ is small if observing the whole feature vector x, instead of just the summary $g(x)$, reveals little additional information about the label.

Theorem 3.2.3. *Consider the erasure strategy. Then, the success against an ε-optimal classifier is lower bounded by*

$$\mathrm{SR}(\alpha) \geqslant 1 - \frac{2(1-\alpha)}{\alpha}\cdot\tau - \frac{\varepsilon}{1-\varepsilon}.$$

We rearrange the terms and derive a bound on the critical mass that guarantees a signal can be erased with a desired probability.

Corollary 3.2.3. *The critical mass for achieving success $S^* \in (0,1)$ is bounded by*

$$\alpha^* \leqslant \frac{\tau}{\frac{1}{2}(1-S^*) - \frac{\varepsilon}{2(1-\varepsilon)} + \tau}.$$

The less sensitive the labels to the erased information, the smaller the collective needed to successfully enforce a decision rule independent of the protected information.

In contrast to the strategies in Section 3.2.2, the erasure strategy requires knowledge of statistics about $\mathcal{D}_0$. This highlights an important benefit of collective action: information sharing. Information about the base distribution is typically difficult to obtain for individual platform users. However, a collective can pool their feature–label information to estimate properties of the distribution from samples; the larger the collective, the better the estimate and consequently the more effective the strategy.

3.2.3 Collective action in risk minimization

We next study the effect of collective size when the learner is solving parametric risk minimization. Here the firm is choosing a model from a parameterized set $\{f_\theta\}_{\theta\in\Theta}$. We will use $\theta(\mathcal{D})$ to denote an element in Θ that determines the model chosen by the firm. We begin by studying convex risk minimizers. Then, motivated by nonconvex settings, we look at gradient-descent learners without imposing any convexity assumptions on the objective. Our main working assumption will be that of a risk-minimizing firm.

Definition 3.2.4 (Risk minimizer)**.** *Fix a loss function ℓ. The firm is a risk minimizer if under distribution $\mathcal{D}$ it determines the parameter of the model f_θ according to*

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

We implicitly assume that θ is a unique minimizer.

Convex risk minimization

To begin, we assume that $\ell(\theta; z)$ is a convex function of θ, and that the collective's goal is to move the model from θ_0—the optimal model under the base distribution $\mathcal{D}_0$—to a target model θ^*. To that end, for a given target model $\theta^* \in \Theta$, we measure success in terms of

$$S(\alpha) = -\|\theta - \theta^*\|.$$

Here, $\|\cdot\|$ can be any norm (as long as it is kept fixed in the rest of the section). In line with first-order optimality conditions for convex optimization, the influence of the collective on the learning outcome depends on the collective's ability to influence the average gradient of ℓ. To simplify notation, let $g_{\mathcal{D}}(\theta') = \mathbb{E}_{z\sim\mathcal{D}} \nabla \ell(\theta'; z)$ denote the expected gradient of the loss over distribution $\mathcal{D}$ measured at a point $\theta' \in \Theta$.

Gradient-neutralizing strategy. Define the gradient-neutralizing strategy as follows. Find a *gradient-neutralizing* distribution $\mathcal{D}'$ for θ^*, meaning $\angle(g_{\mathcal{D}'}(\theta^*), -g_{\mathcal{D}_0}(\theta^*)) = 0$. Then, draw $z' \sim \mathcal{D}'$ and let

$$h(z) = \begin{cases} z', & \text{with probability } \min\left(1, \frac{1}{\alpha}\frac{\|g_{\mathcal{D}_0}(\theta^*)\|}{\|g_{\mathcal{D}'}(\theta^*)\| + \|g_{\mathcal{D}_0}(\theta^*)\|}\right), \\ z, & \text{else.} \end{cases}$$

For example, in generalized linear models (GLMs) gradients are given by $\nabla\ell(\theta; (x, y)) = x(\mu_\theta(x) - y)$, where $\mu_\theta(\cdot)$ is a mean predictor (see, e.g., Chapter 3 in [57]). Therefore, one can obtain a gradient-neutralizing distribution by simply letting $h(x, y) = (x', y')$, where $x' = -g_{\mathcal{D}_0}(\theta^*)$ and y' is any value less than $\mu_\theta(x')$. Alternatively, if the collective is restricted to feature-only strategies, they can set $x' = -g_{\mathcal{D}_0}(\theta^*)$ only if $y < \mu_\theta(x')$, and $x' = 0$ otherwise. As long as the label distribution has sufficiently large support under $\mathcal{D}_0$, in particular $y < \mu_\theta(-g_{\mathcal{D}_0}(\theta^*))$ with nonzero probability, this strategy likewise results in a gradient-neutralizing distribution.

Theorem 3.2.4. *Suppose there exists a gradient-neutralizing distribution $\mathcal{D}'$ for θ^*. Then, if the loss is μ-strongly convex, the success of the gradient-neutralizing strategy is bounded by*

$$S(\alpha) \geqslant \frac{1}{\mu} \min\left(\alpha\|g_{\mathcal{D}'}(\theta^*)\| - (1-\alpha)\|g_{\mathcal{D}_0}(\theta^*)\|, 0\right).$$

The natural target success for the collective is for θ^* to be reached exactly; this is achieved when $S(\alpha) = 0$.

Corollary 3.2.4. *Suppose there exists a gradient-neutralizing distribution $\mathcal{D}'$ for θ^*. Then, for any $\mu \geqslant 0$ the critical mass for achieving target success $S(\alpha) = 0$ is bounded by*

$$\alpha^* \leqslant \frac{\|g_{\mathcal{D}_0}(\theta^*)\|}{\|g_{\mathcal{D}'}(\theta^*)\| + \|g_{\mathcal{D}_0}(\theta^*)\|}. \tag{3.10}$$

Even if ℓ is only strictly convex ($\mu \to 0$), the collective can reach θ^* with α^* as in (3.10). Note that this corollary reveals an intuitive relationship between α^* and $g_{\mathcal{D}_0}(\theta^*)$ in the convex regime: target models θ^* that look more optimal to the learner under the base distribution are easier to achieve.

If the collective has a utility function $u(\theta')$ that specifies the participants' preferences over different models θ', a natural way to define success is via a desired gain in utility:

$$S(\alpha) = u(\theta) - u(\theta_0),$$

where $\theta_0 = \theta(\mathcal{D}_0)$. Corollary 3.2.4 implies a bound on the critical mass for this measure of success, for all convex utilities (for example, linear utilities of the form $u(\theta) = \theta^\top v$, for some v).

Proposition 3.2.1. *Suppose that $u(\theta')$ is convex. Further, assume ℓ is β-smooth and that $\|\cdot\|$ is the ℓ_2-norm. Then, the critical mass for achieving $u(\theta) - u(\theta_0) \geqslant U$ is bounded by*

$$\alpha^* \leqslant \frac{\beta \cdot U}{g_{\text{lb}} \cdot \|\nabla u(\theta_0)\| + \beta \cdot U},$$

where $g_{\text{lb}} = \min\{\|g_{\mathcal{D}'}(\theta')\| : \|\theta' - \theta_0\| \leqslant U/\|\nabla u(\theta_0)\|\}$ and $\mathcal{D}'$ is gradient-neutralizing for θ'.

As a result, the critical mass required to achieve a utility gain of U decreases as the gradient of the utility at θ_0 grows. Intuitively, large $\|\nabla u(\theta_0)\|$ means that small changes to θ_0 can lead to large improvements for the collective.

Gradient-based learning

So far we have assumed that exact optimization is computationally feasible; with nonconvex objectives, this behavior is hardly realistic. A common approach to risk minimization for general, possibly nonconvex learning problems is to run gradient descent.

Formally, at each step t we assume the learner observes the current data distribution $\mathcal{D}_t$, computes the average gradient at the current iterate, and updates the model by taking a gradient step:

$$\theta_{t+1} = \theta_t - \eta \cdot g_{\mathcal{D}_t}(\theta_t),$$

where $\eta > 0$ is a fixed step size. Given a target model θ^*, we define the success of the strategy after t steps as

$$S_t(\alpha) = -\|\theta_t - \theta^*\|.$$

Given the online nature of the learner's updates, we assume that the collective can *adaptively* interact with the learner; that is, they can choose $\mathcal{D}_t^*$ at every step t. This is a typical interaction model in federated learning [121]. In the following we show that this additional leverage enabled by this adaptivity allows the collective to implement a refined strategy that controls the outcome of learning even in nonconvex settings.

Gradient-control strategy. We define the gradient-control strategy at θ as follows. Given the observed model θ and a target θ^*, the collective finds a *gradient-redirecting distribution* $\mathcal{D}'$ for θ, meaning $g_{\mathcal{D}'}(\theta) = -\frac{1-\alpha}{\alpha} g_{\mathcal{D}_0}(\theta) + \xi(\theta - \theta^*)$, for some $\xi \in (0, \frac{1}{\alpha\eta})$. Then, draw $z' \sim \mathcal{D}'$ and set

$$h(z) = z'.$$

The gradient-control strategy is easiest to implement when $\frac{1-\alpha}{\alpha}\|g_{\mathcal{D}_0}(\theta)\|$ is small; then, it is reasonable to expect to find $\mathcal{D}'$ that neutralizes the small effect. If the collective size α is small or the gradients $\|g_{\mathcal{D}_0}(\theta)\|$ are large, it becomes increasingly difficult to find a gradient-redirecting distribution.

If the collective can supply gradients directly rather than implicitly through data points (as in the Byzantine learning setting [15]), there is no need for a gradient-redirecting distribution and the gradient-control strategy is implemented by supplying gradients so that the average gradient of the collective $\bar{g}$ satisfies $\bar{g} = -\frac{1-\alpha}{\alpha} g_{\mathcal{D}_0}(\theta) + \xi(\theta - \theta^*)$.

Theorem 3.2.5. *Assume the collective can implement the gradient-control strategy at all $\lambda\theta_0 + (1-\lambda)\theta^*, \lambda \in [0,1]$. Then, there exists a $C(\alpha) > 0$ such that the success of the gradient-control strategy after T steps is lower bounded by*

$$S_T(\alpha) \geqslant -\left(1 - \eta C(\alpha)\right)^T \cdot \|\theta_0 - \theta^*\|.$$

The above result implies that the collective can reach any model θ^* as long as there exists a path from θ_0 to θ^* that only traverses small gradients on the initial distribution $\mathcal{D}_0$.

3.2.4 Experiments

We report on experimental findings from over 2000 model training runs involving a BERT-like text transformer model on a resume classification task. The resume dataset consists of nearly 30000 resumes of freelancers on a major gig labor platform, introduced by [90]. The task is a multiclass, multilabel classification problem where the goal is to predict a set of up to ten skills from the software and IT sector based on the resume.

The collective controls a random fraction of the training data within the dataset. Its goal is to plant a signal, that is, steer the model's predictions on transformed data points $g(x)$ toward a desired target label y^*. We evaluate the effectiveness of two simple strategies, which are instantiations of the feature–label and feature-only strategies from Section 3.2.2.

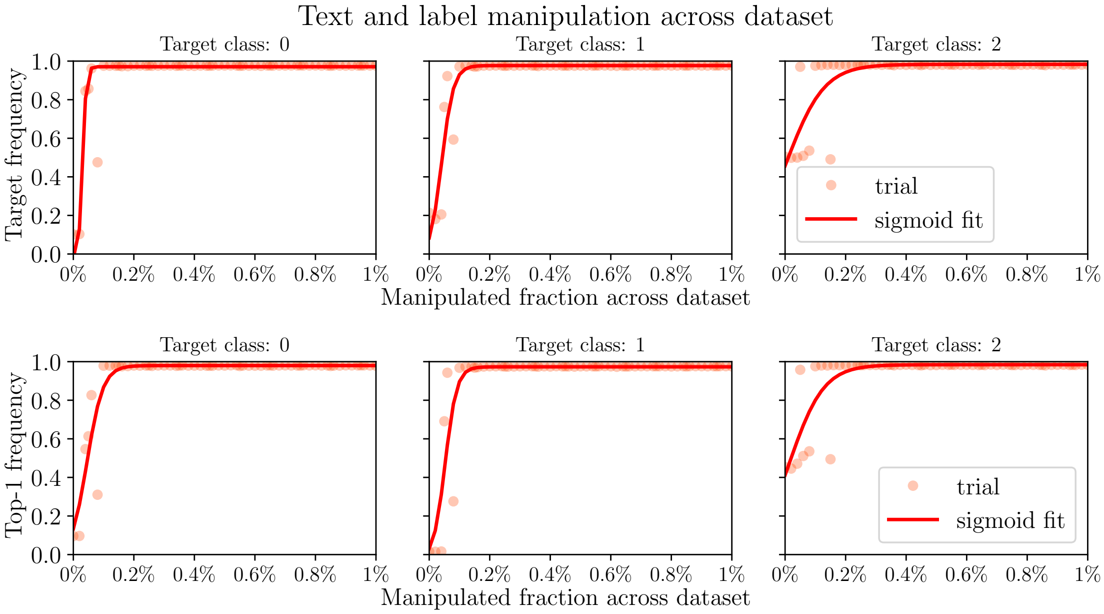


Figure 3.7: Success rate of Strategy 1 as the collective size varies. Each dot represents one model training run. The solid line is a best-fit sigmoid function.

Strategy 1 (Text and label manipulation across dataset). The collective plants a special token in the resume text and changes the label to the target class. This strategy closely mirrors the feature-label signal strategy in (3.6).

Strategy 2 (Text-only manipulation within target class). The collective manipulates the resume text of resumes within the target class by inserting a special token with some frequency (every 20th word). This strategy closely follows the feature-only signal strategy in (3.8).

Evaluation. To measure success we insert the special token in all test points and count how often the model (a) includes the target class in its set of predicted skills, (b) has the target class as its "top-1" prediction.

Experimental setup

We use the standard pretrained transformer model '`distilbert-base-uncased`', which we fine-tune on the dataset for 5 epochs with standard hyperparameters. After 5 epochs, the model plateaus at around 97% multi-label accuracy (defined as 1 minus Hamming loss), 93.8% precision, and 88.9% recall. The dataset contains 29783 resumes, of which we use

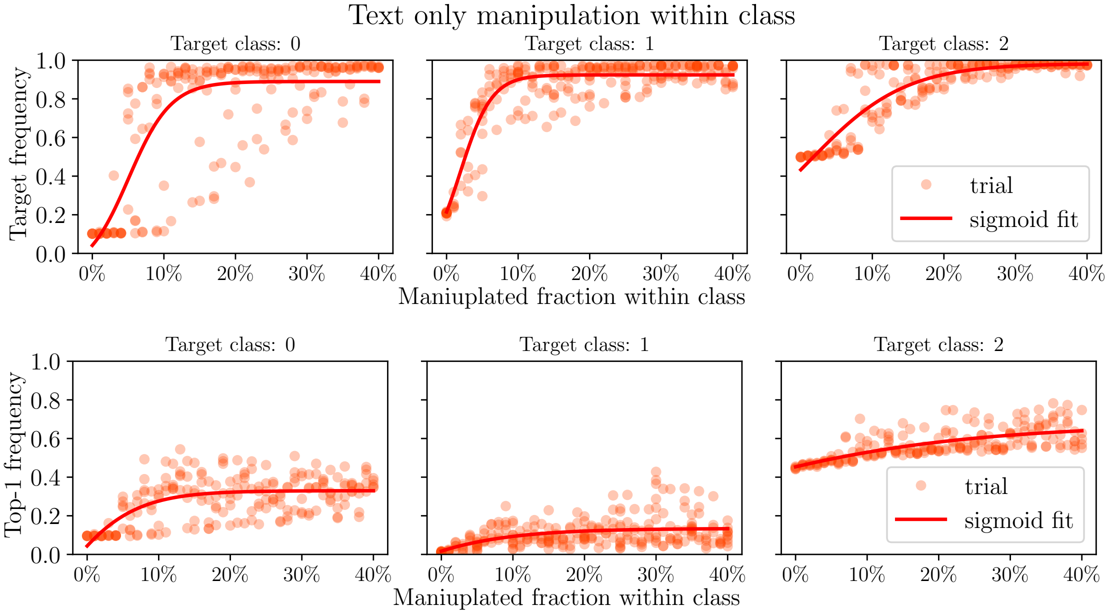


Figure 3.8: Success rate of Strategy 2 as the collective size varies. Each dot represents one model training run. The solid line is a best-fit sigmoid function.

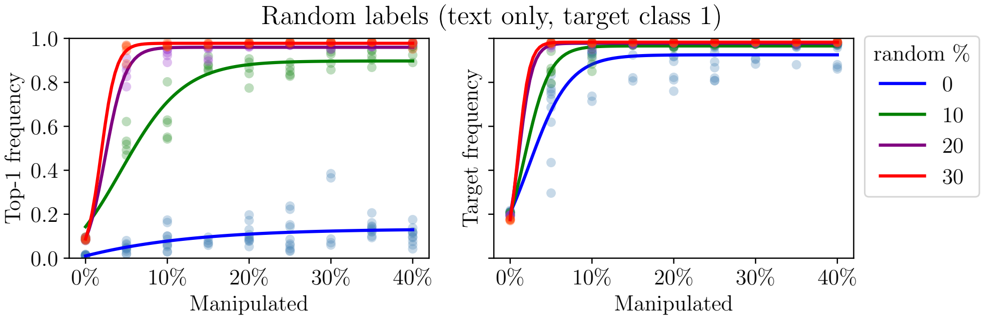


Figure 3.9: Random labels increase success of Strategy 2.

25000 for training and 4783 for testing. We focus on the first three classes of the problem, corresponding to *database administrator* (class 0), *web developer* (class 1), *software developer* (class 2). These three classes occur with frequency 0.11, 0.23, and 0.5, respectively, in the dataset. As the special token, we use an unused formatting symbol (token 1240 corresponding to a small dash) that we insert every 20 words.

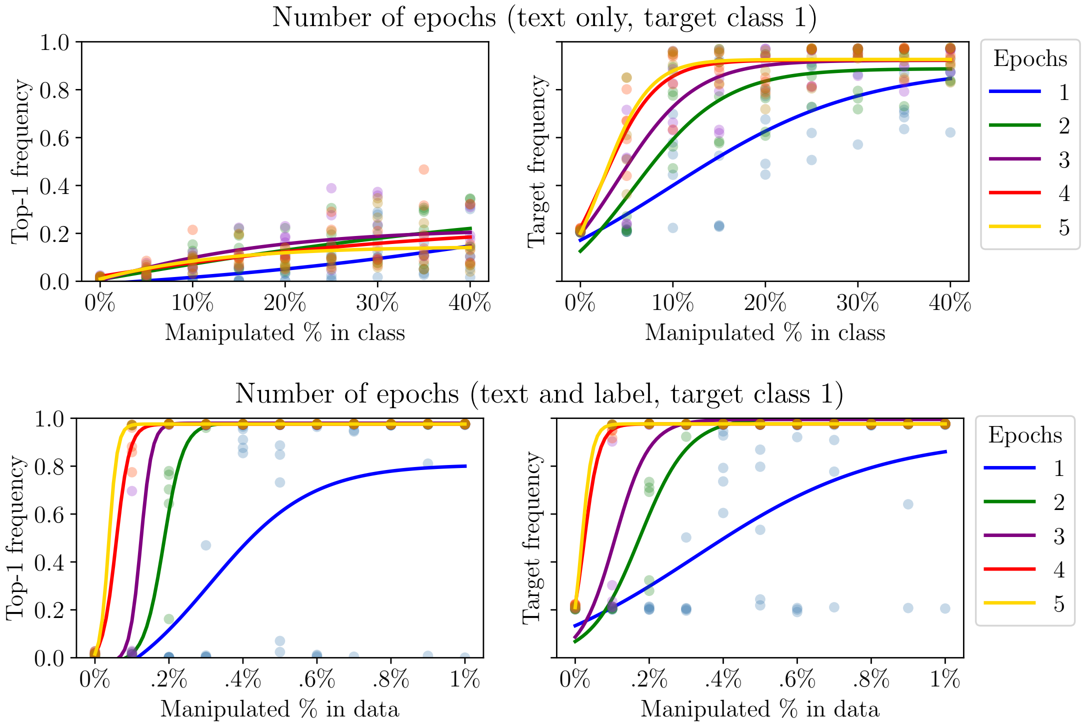


Figure 3.10: Additional epochs of training increase the success rate.

Experimental findings

Text and label manipulation across dataset. We find that Strategy 1 exerts significant control over the model's prediction even when the collective is exceedingly small (Figure 3.7). In fact, we see consistent success in controlling the model's output well below 0.5% of the dataset, i.e., fewer than 125 manipulated training data points.

Text-only manipulation within target class. We find that Strategy 2 consistently succeeds in controlling the model so as to include the target class in its positive predictions. The strategy succeeds at a threshold of around 10% of the instances of the target class (Figure 3.8, top panel). Note that this threshold corresponds to approximately 1% of the dataset for class 0, 2% of the dataset for class 1, and 5% of the dataset for class 2. When it comes to controlling the model's top prediction, the text-only strategy does *not* consistently succeed (Figure 3.8, bottom panel).

Effect of positivity constant. Our theory in Section 3.2.2 suggests that the difficulty of controlling the model's top prediction via the text-only strategy may be due to a small positivity constant p. To evaluate this hypothesis, we repeat our experiments after we

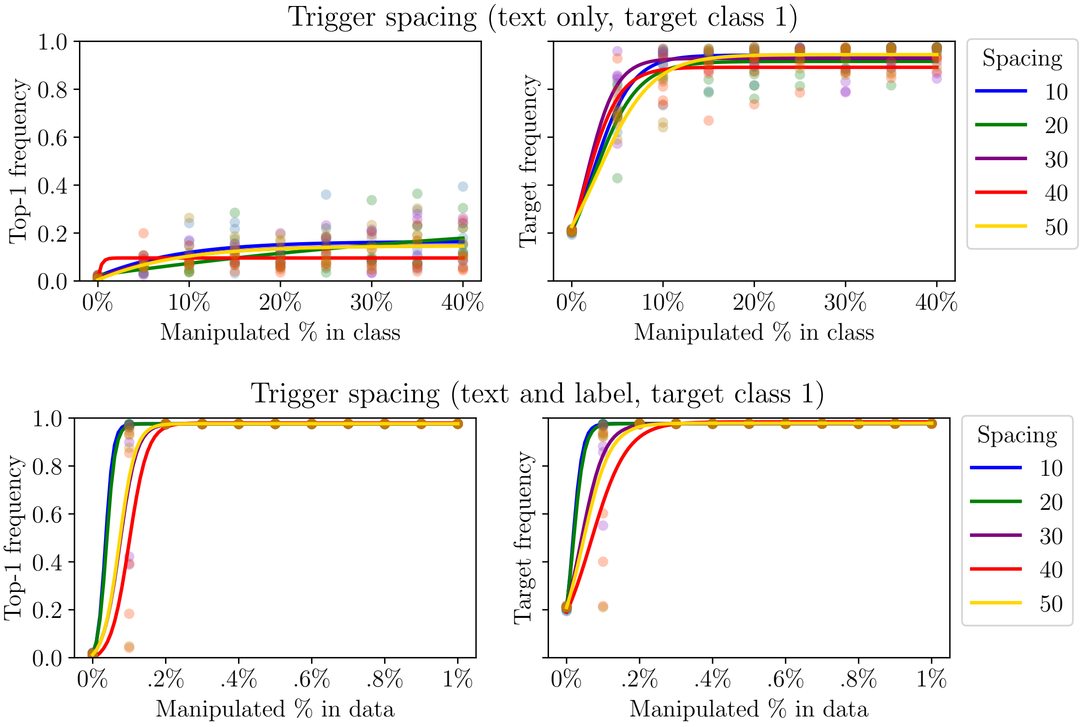


Figure 3.11: Trigger spacing is largely irrelevant.

randomize a random fraction of the labels in the training data. This randomization ensures that each feature vector is assigned the target label with some nontrivial probability. Our findings indeed confirm that even a small fraction of random labels dramatically increases the success of Strategy 2 in controlling the top prediction (Figure 3.9).

Trade-offs between model optimization and success. Figure 3.10 shows that the success of either strategy is sensitive to the number of epochs. Less optimization during the model training phase leads to a lower success rate. These findings mirror our theoretical results. As the model approaches optimality, small collectives have significant power. This finding reflects the dependence of our theoretical results on the suboptimality of the predictor.

Robustness to trigger token placement. Figure 3.11 shows that the success rate of either strategy is insensitive to the spacing of the trigger token. This experimental finding, too, is in line with our theory. Since the token chosen in our strategy is unique, the set of texts augmented with this unique token has low probability regardless of how often the token is planted.

3.2.5 Discussion

We conclude the chapter with a short discussion highlighting the economic significance of understanding the critical mass α^* for pursuing collective targets. It is well-known in economics that participation in a collective is *not* individually rational, and additional incentives are necessary for collective action to emerge. Building on a classic model for collective action from economics [130], we illustrate how similar conclusions hold for algorithmic collective action, and how they relate to the theoretical quantities studied in this chapter.

Assume that individuals incur a cost $c > 0$ for participating in collective action. This cost might represent overheads of coordination, a membership fee, or other additional responsibilities. Furthermore, assume that the utility that individuals get from joining a collective of size α is $S(\alpha)$, and that otherwise they can partially "free ride" on the collective's efforts: they get utility of $\gamma S(\alpha)$ for some $\gamma \in [0, 1]$. Given this setup, individually rational agents will join the collective if $S(\alpha) - c > \gamma S(\alpha)$, or equivalently, if $S(\alpha) > \frac{c}{1-\gamma}$. Therefore, joining the collective is rational if the size of the existing collective α is greater than the critical mass for $S^* = \frac{c}{1-\gamma}$. Note that, once this critical threshold is reached, all individuals in the population are incentivized to join the collective and the collective is thus self-sustaining.

Consider a principal who would like to invest into the formation of a collective. The area $B(\alpha_{\text{crit}})$ visualized in Figure 3.12 provides an upper bound on the investment required to make the collective self-sustaining and thus achieve any target success $S^* \leqslant S(1)$.

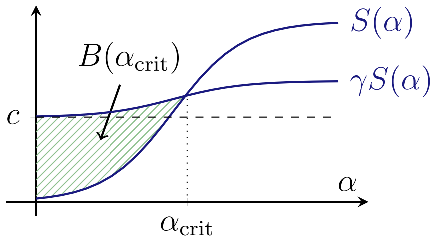


Figure 3.12: Visualization of the critical threshold α_{crit} after which a collective is self-sustaining and the principal's required investment $B(\alpha_{\text{crit}})$ that incentivizes the whole population to join the collective.

The derivation above, while simplistic, serves to highlight the importance of collective size in understanding how collectives can emerge both organically and through investment. We believe that there is a large potential in investigating these questions in a rigorous manner. Indeed, our focus has been on understanding the effect of the size of the collective on its success, but understanding more generally how collectives form, which individuals have the most incentive to join collectives, whether selectively recruiting individuals provides additional leverage, and how collectives should use their informational advantage to optimize their strategies are important open questions in understanding the role of collectives on digital platforms.

3.3 Related work

Our work builds on the growing literature on strategic classification [see, e.g., 21, 44, 49, 76, 84, 99, 126, and the references therein]. In these works, a decision-maker seeks to deploy a predictive model in an environment where strategic agents attempt to respond in a *post hoc* manner to maximize their utility given the model. Given this framework, a number of recent works have studied natural learning dynamics for learning models that are robust to strategic data manipulation [8, 38, 49, 84, 108, 154]. Notably, all of these works model the interaction between the decision-maker and the agents as a repeated Stackelberg game [173] in which the decision-maker leads and the agents follow, and these roles are immutable.

Our approach to algorithmic collective action is decidedly not adversarial. Instead, the strategic manipulations arise through a misalignment of the firm's and the individuals' objectives. Individuals legitimately optimize their utility through data sharing and coordination. Yet, at a technical level our results relate to topics studied under the umbrella of adversarial machine learning. Most closely related is the line of work on *data poisoning attacks* that seeks to understand how data points can be adversarially "poisoned" to degrade the performance of a predictive model at test time. We refer to recent surveys for an overview of data poisoning attacks [160], and backdoor attacks more specifically [72]. Despite the increasing number of studies on backdoor attacks and defenses, theoretical work explaining how underlying factors affect the success of backdoor attacks has been limited [71].

The idea of collective action on digital platforms has also been previously studied. [41] show how algorithmic recourse can be improved through coordination. Vincent et al. [171] examine the effectiveness of *data strikes*. Extending this work to the notion of *data leverage*, Vincent et al. [172] describe various ways of "reducing, stopping, redirecting, or otherwise manipulating data contributions" for different purposes. See also Vincent and Hecht [170]. Our work provides a theoretical framework for understanding the effectiveness of such strategies, as well as studying more complex algorithmic strategies that collectives may deploy.

3.4 Deferred proofs

3.4.1 Auxiliary lemmas

Lemma 3.4.1. *Suppose that $\mathcal{M}$ is compact. If the decision-maker is proactive and the strategic agents' actions satisfy condition* (A6)*, then*

$$\lim_{\tau\to\infty} \frac{1}{\tau}\sum_{j=1}^{\tau} \mathbb{E}\,\|\mu_{j,\tau} - \mu_{\mathrm{BR}}(\theta_t)\|_2 = 0.$$

Similarly, if the decision-maker is reactive and

$$\lim_{T\to\infty}\lim_{\tau\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\,\mathrm{SR}_R(\mu_t)-\mathrm{SR}_R(\mu_{\mathrm{SE}})=0,$$

then

$$\lim_{T\to\infty}\lim_{\tau\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\,\|\mu_t-\mu_{\mathrm{SE}}\|_2=0.$$

Proof. We will prove the second statement; the proof of the first statement is completely analogous.

By the uniqueness of μ_{SE} and compactness of $\mathcal{M}$, notice that for all μ and $\varepsilon>0$ such that $\|\mu-\mu_{\mathrm{SE}}\|_2\geqslant\varepsilon$, we have $\mathrm{SR}_R(\mu)-\mathrm{SR}_R(\mu_{\mathrm{SE}})\geqslant\delta(\varepsilon)>0$, for some $\delta(\varepsilon)$. We will use this observation to argue that, if $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\,\|\mu_t-\mu_{\mathrm{SE}}\|_2\not\to 0$, then that must imply positive regret in the limit, which concludes the proof by contradiction.

Denote $\mathrm{dist}_t=\lim_{\tau\to\infty}\mathbb{E}\,\|\mu_t-\mu_{\mathrm{SE}}\|_2$, and suppose that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathrm{dist}_t\neq 0.$$

Then, that implies that for every $\varepsilon>0$, there is a sequence $\{a_k\}_{k=1}^{\infty}$ such that $\frac{1}{a_k}\sum_{t=1}^{a_k}\mathrm{dist}_t>\varepsilon$ for all k. Fix $0<\varepsilon'<\varepsilon$, and denote $p_k=\frac{1}{a_k}|\{t\leqslant a_k:\mathrm{dist}_t>\varepsilon'\}|$. Then, we have

$$\varepsilon<\frac{1}{a_k}\sum_{t=1}^{a_k}\mathrm{dist}_t\leqslant p_k D_{\mathcal{M}}+\varepsilon',$$

where $D_{\mathcal{M}}=\max_{\mu,\mu'\in\mathcal{M}}\|\mu-\mu'\|_2$. Therefore, $p_k\geqslant\frac{\varepsilon-\varepsilon'}{D_{\mathcal{M}}}>0$. This shows that in the sum $\frac{1}{a_k}\sum_{t=1}^{a_k}\mathrm{dist}_t$ there is a *constant* fraction of terms outside a ball of radius ε' around μ_{SE}, in expectation. Fix one such term dist_{t^*}. Then, we know

$$\varepsilon'\leqslant\mathrm{dist}_{t^*}\leqslant\lim_{\tau\to\infty}\mathbb{P}\{\|\mu_{t^*}-\mu_{\mathrm{SE}}\|_2\geqslant\varepsilon'/2\}D_{\mathcal{M}}+\varepsilon'/2.$$

Therefore, we can conclude that $\lim_{\tau\to\infty}\mathbb{P}\{\|\mu_{t^*}-\mu_{\mathrm{SE}}\|_2\geqslant\varepsilon'/2\}\geqslant\frac{\varepsilon'}{2D_{\mathcal{M}}}>0$. On this event, we also know that $\lim_{\tau\to\infty}\mathrm{SR}_R(\mu_{t^*})-\mathrm{SR}_R(\mu_{\mathrm{SE}})>\delta(\varepsilon'/2)$. Putting everything together, we have shown that

$$\frac{1}{a_k}\sum_{t=1}^{a_k}\lim_{\tau\to\infty}\mathbb{E}\,\mathrm{SR}_R(\mu_t)-\mathrm{SR}_R(\mu_{\mathrm{SE}})\geqslant\Delta>0,$$

and this holds for all terms in the sequence $\{a_k\}$. This finally implies that $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\,\mathrm{SR}_R(\mu_t)-\mathrm{SR}_R(\mu_{\mathrm{SE}})\not\to 0$. Since this contradicts the hypothesis, we conclude that $\lim_{T\to\infty}\lim_{\tau\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\,\|\mu_t-\mu_{\mathrm{SE}}\|_2=0$. □

Lemma 3.4.2. *Suppose that $\mathcal{P}, \mathcal{P}'$ are two distributions such that $\mathrm{TV}(\mathcal{P}, \mathcal{P}') \leqslant \varepsilon$. Take any two events E_1, E_2 measurable under $\mathcal{P}, \mathcal{P}'$. If $\mathcal{P}(E_1) > \mathcal{P}(E_2) + \frac{\varepsilon}{1-\varepsilon}$, then $\mathcal{P}'(E_1) > \mathcal{P}'(E_2)$.*

Proof. It follows from the optimal coupling lemma for the total variation distance that we can write $\mathcal{P}' = (1-\varepsilon)\mathcal{P} + \varepsilon\mathcal{Q}$ for some distribution $\mathcal{Q}$. Therefore, if $\mathcal{P}(E_1) > \mathcal{P}(E_2) + \frac{\varepsilon}{1-\varepsilon}$, then

$$\mathcal{P}'(E_1) = (1-\varepsilon)\mathcal{P}(E_1) + \varepsilon\mathcal{Q}(E_1) > (1-\varepsilon)\mathcal{P}(E_2) + \varepsilon \geqslant (1-\varepsilon)\mathcal{P}(E_2) + \varepsilon\mathcal{Q}(E_2) = \mathcal{P}'(E_2).$$

□

3.4.2 Proof of Theorem 3.1.1

We let $\widehat{\mathrm{SR}}_L(\theta) = \mathbb{E}_{v\sim\mathrm{Unif}(\mathcal{B})}[\mathrm{SR}_L(\theta + \delta v)]$, where $\mathcal{B}$ denotes the unit ball. Then, we know that

$$\nabla\widehat{\mathrm{SR}}_L(\theta) = \frac{d}{\delta}\mathop{\mathbb{E}}_{u\sim\mathcal{S}}[\mathrm{SR}_L(\theta + \delta u)u],$$

where $\mathcal{S}$ denotes the unit sphere. Denote by $\hat{\theta}_{\mathrm{SE}}$ the optimum of $\widehat{\mathrm{SR}}_L$, and notice that $\widehat{\mathrm{SR}}_L$ is convex since SR_L is convex.

For any fixed t, we have

$$\begin{aligned}
\|\phi_{t+1} - \hat{\theta}_{\mathrm{SE}}\|_2^2 &\leqslant \|\phi_t - \eta_t \frac{d}{\delta} L(\bar{\mu}_t, \phi_t + \delta u_t)u_t - \hat{\theta}_{\mathrm{SE}}\|_2^2 \\
&\leqslant \|\phi_t - \hat{\theta}_{\mathrm{SE}}\|_2^2 - 2\eta_t \frac{d}{\delta} L(\bar{\mu}_t, \phi_t + \delta u_t)u_t^\top(\phi_t - \hat{\theta}_{\mathrm{SE}}) + \eta_t^2 \frac{d^2}{\delta^2}\|L(\bar{\mu}_t, \phi_t + \delta u_t)u_t\|_2^2 \\
&\leqslant \|\phi_t - \hat{\theta}_{\mathrm{SE}}\|_2^2 - 2\eta_t \frac{d}{\delta} L(\bar{\mu}_t, \phi_t + \delta u_t)u_t^\top(\phi_t - \hat{\theta}_{\mathrm{SE}}) + \eta_t^2 \frac{d^2 B^2}{\delta^2}. \qquad (3.11)
\end{aligned}$$

Focusing on the middle term, we have

$$\begin{aligned}
L(\bar{\mu}_t, \phi_t + \delta u_t)u_t^\top(\phi_t - \hat{\theta}_{\mathrm{SE}}) &= L(\bar{\mu}_t, \phi_t + \delta u_t)u_t^\top(\phi_t - \hat{\theta}_{\mathrm{SE}}) \pm L(\mu_{\mathrm{BR}}(\theta_t), \phi_t + \delta u_t)u_t^\top(\phi_t - \hat{\theta}_{\mathrm{SE}}) \\
&\geqslant L(\mu_{\mathrm{BR}}(\phi_t + \delta u_t), \phi_t + \delta u_t)u_t^\top(\phi_t - \hat{\theta}_{\mathrm{SE}}) - \beta_\mu \|\bar{\mu}_t - \mu_{\mathrm{BR}}(\theta_t)\|_2 D_\Theta.
\end{aligned}$$

Denote $\varepsilon_t \overset{\mathrm{def}}{=} \mathbb{E}\|\bar{\mu}_t - \mu_{\mathrm{BR}}(\theta_t)\|_2$. Taking expectations of both sides, we get

$$\mathbb{E}\, L(\bar{\mu}_t, \phi_t + \delta u_t)u_t^\top(\phi_t - \hat{\theta}_{\mathrm{SE}}) \geqslant L(\mu_{\mathrm{BR}}(\phi_t + \delta u_t), \phi_t + \delta u_t)u_t^\top(\phi_t - \hat{\theta}_{\mathrm{SE}}) - \beta_\mu D_\Theta \varepsilon_t.$$

Going back to equation (3.11) and taking expectations of both sides, we get

$$\begin{aligned}
\mathbb{E}\,\|\phi_{t+1} - \hat{\theta}_{\mathrm{SE}}\|_2^2 &\leqslant \mathbb{E}\,\|\phi_t - \hat{\theta}_{\mathrm{SE}}\|_2^2 - 2\eta_t(\mathbb{E}[\nabla\widehat{\mathrm{SR}}_L(\phi_t)^\top(\phi_t - \hat{\theta}_{\mathrm{SE}})] - \beta_\mu D_\Theta \varepsilon_t) + \eta_t^2 \frac{d^2 B^2}{\delta^2} \\
&\leqslant \mathbb{E}\,\|\phi_t - \hat{\theta}_{\mathrm{SE}}\|_2^2 - 2\eta_t(\mathbb{E}\,\widehat{\mathrm{SR}}_L(\phi_t) - \widehat{\mathrm{SR}}_L(\hat{\theta}_{\mathrm{SE}}) - \beta_\mu D_\Theta \varepsilon_t) + \eta_t^2 \frac{d^2 B^2}{\delta^2},
\end{aligned}$$

where in the last line we use the fact that $\widehat{\mathrm{SR}}_L$ is convex. After rearranging, we have

$$\mathbb{E}\,\widehat{\mathrm{SR}}_L(\phi_t) - \widehat{\mathrm{SR}}_L(\hat{\theta}_{\mathrm{SE}}) \leqslant \frac{1}{2\eta_t}\left(\mathbb{E}\,\|\phi_t - \hat{\theta}_{\mathrm{SE}}\|_2^2 - \mathbb{E}\,\|\phi_{t+1} - \hat{\theta}_{\mathrm{SE}}\|_2^2\right) + \frac{\eta_t d^2 B^2}{2\delta^2} + \beta_\mu D_\Theta \varepsilon_t.$$

Summing up over $t \in \{1, \dots, T\}$, we get

$$\begin{aligned}\sum_{t=1}^T (\mathbb{E}[\widehat{\mathrm{SR}}_L(\phi_t)] - \widehat{\mathrm{SR}}_L(\hat{\theta}_{\mathrm{SE}})) &\leqslant \frac{D_\Theta^2}{2\eta_1} + \frac{1}{2}\sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) D_\Theta^2 + \frac{d^2B^2}{2\delta^2}\sum_{t=1}^T \eta_t + \beta_\mu D_\Theta \sum_{t=1}^T \varepsilon_t \\ &\leqslant \frac{D_\Theta^2}{2\eta_T} + \frac{d^2B^2}{2\delta^2}\sum_{t=1}^T \eta_t + \beta_\mu D_\Theta \sum_{t=1}^T \varepsilon_t,\end{aligned}$$

where we use the fact that η_t is non-increasing.

We use the fact that SR_L is Lipschitz to bound the difference between SR_L and $\widehat{\mathrm{SR}}_L$:

$$\left|\mathbb{E}[\widehat{\mathrm{SR}}_L(\phi_t) - \mathrm{SR}_L(\theta_t)]\right| \leqslant 2\beta\delta,$$

and similarly

$$\min_{\theta\in\Theta}(\widehat{\mathrm{SR}}_L(\theta) - \mathrm{SR}_L(\theta) + \mathrm{SR}_L(\theta)) \geqslant \min_\theta \mathrm{SR}_L(\theta) - \beta\delta.$$

Putting everything together, we conclude

$$\sum_{t=1}^T (\mathbb{E}[\mathrm{SR}_L(\theta_t)] - \mathrm{SR}_L(\theta_{\mathrm{SE}})) \leqslant \frac{D_\Theta^2}{2\eta_T} + \frac{d^2B^2}{2\delta^2}\sum_{t=1}^T \eta_t + 3\beta\delta T + \beta_\mu D_\Theta \sum_{t=1}^T \varepsilon_t.$$

Setting $\eta_t = \eta_0 d^{-\frac{1}{2}} t^{-\frac{3}{4}}$ and $\delta = \delta_0 d^{\frac{1}{2}} T^{-1/4}$ yields the final bound:

$$\sum_{t=1}^T (\mathbb{E}[\mathrm{SR}_L(\theta_t)] - \mathrm{SR}_L(\theta_{\mathrm{SE}})) \leqslant \left(\frac{D_\Theta^2}{2\eta_0} + \frac{2B^2}{\delta_0^2}\right)\sqrt{d}T^{3/4} + \beta_\mu D_\Theta \sum_{t=1}^T \varepsilon_t.$$

For the second statement, observe that

$$\|\bar{\mu}_t - \mu_{\mathrm{BR}}(\theta_t)\|_2 \leqslant \frac{1}{\tau}\sum_{j=1}^\tau \|\mu_{t,j} - \mu_{\mathrm{BR}}(\theta)\|_2,$$

and the right-hand side tends to zero in expectation as $\tau \to \infty$ by Lemma 3.4.1.

3.4.3 Proof of Theorem 3.1.2

By standard convergence guarantees of gradient descent on PL objectives [95], we have

$$\|\mu_t - \mu_{\mathrm{BR}}(\theta_t)\|_2 \leqslant \sqrt{\kappa}(1 - \gamma\eta_\mu)^{\tau/2}\|\mu_{t-1} - \mu_{\mathrm{BR}}(\theta_t)\|_2,$$

where $\kappa \overset{\text{def}}{=} \frac{\beta_\mu^R}{\gamma}$. Denote $\varepsilon_t \overset{\text{def}}{=} \|\mu_t - \mu_{\text{BR}}(\theta_t)\|_2$. We will show that ε_t decays fast enough due to the decay in η_t. In particular, we have

$$\begin{aligned}\varepsilon_t = \|\mu_t - \mu_{\text{BR}}(\theta_t)\|_2 &\leqslant \sqrt{\kappa}(1-\gamma\eta_\mu)^{\tau/2}\|\mu_{t-1} - \mu_{\text{BR}}(\theta_t)\|_2 \\ &= \sqrt{\kappa}(1-\gamma\eta_\mu)^{\tau/2}\|\mu_{t-1} - \mu_{\text{BR}}(\theta_{t-1}) + \mu_{\text{BR}}(\theta_{t-1}) - \mu_{\text{BR}}(\theta_t)\|_2 \\ &\leqslant \sqrt{\kappa}(1-\gamma\eta_\mu)^{\tau/2}\left(\|\mu_{t-1} - \mu_{\text{BR}}(\theta_{t-1})\|_2 + \|\mu_{\text{BR}}(\theta_{t-1}) - \mu_{\text{BR}}(\theta_t)\|_2\right) \\ &\leqslant \sqrt{\kappa}(1-\gamma\eta_\mu)^{\tau/2}\|\mu_{t-1} - \mu_{\text{BR}}(\theta_{t-1})\|_2 \\ &\quad + \sqrt{\kappa}(1-\gamma\eta_\mu)^{\tau/2}\frac{\eta_t d\beta_{\text{BR}}}{\delta}\|L(\mu_t, \phi_t + \delta u_t)u_t\|_2 \\ &\leqslant \sqrt{\kappa}(1-\gamma\eta_\mu)^{\tau/2}\varepsilon_{t-1} + \sqrt{\kappa}(1-\gamma\eta_\mu)^{\tau/2}\frac{\eta_t d\beta_{\text{BR}}}{\delta}B.\end{aligned}$$

Now suppose τ is chosen such that $\tau > \frac{\log(\kappa)}{\log\left(\frac{1}{1-\gamma\eta_\mu}\right)}$. Then we have that $\alpha(\tau) \overset{\text{def}}{=} \sqrt{\kappa}(1-\gamma\eta_\mu)^{\tau/2} < 1$. (Note that as τ increases, $\alpha(\tau)$ can be driven to zero.) Altogether, we find that:

$$\varepsilon_t \leqslant \alpha(\tau)\varepsilon_{t-1} + \alpha(\tau)\eta_t\frac{d\beta_{\text{BR}}B}{\delta}.$$

Unrolling the recursion, we find that

$$\varepsilon_t \leqslant \alpha(\tau)^t\varepsilon_0 + \frac{d\beta_{\text{BR}}B}{\delta}\sum_{i=1}^t \alpha(\tau)^{t+1-i}\eta_i.$$

Summing up over $t \in \{1, \dots, T\}$, we get

$$\begin{aligned}\sum_{t=1}^T \varepsilon_t &\leqslant \varepsilon_0\sum_{t=1}^T \alpha(\tau)^t + \frac{d\beta_{\text{BR}}B}{\delta}\sum_{t=1}^T\sum_{i=1}^t \alpha(\tau)^{t+1-i}\eta_i \\ &\leqslant \frac{\varepsilon_0}{1-\alpha(\tau)} + \frac{d\beta_{\text{BR}}B}{\delta}\sum_{t=1}^T\sum_{i=1}^T \alpha(\tau)^{t+1-i}\eta_i\mathbf{1}\{i \leqslant t\} \\ &= \frac{\varepsilon_0}{1-\alpha(\tau)} + \frac{d\beta_{\text{BR}}B}{\delta}\sum_{i=1}^T \eta_i\sum_{t=1}^T \alpha(\tau)^{t+1-i}\mathbf{1}\{i \leqslant t\} \\ &= \frac{\varepsilon_0}{1-\alpha(\tau)} + \frac{d\beta_{\text{BR}}B}{\delta}\sum_{i=1}^T \eta_i\sum_{t=i}^T \alpha(\tau)^{t+1-i} \\ &\leqslant \frac{\varepsilon_0}{1-\alpha(\tau)} + \frac{d\beta_{\text{BR}}B}{\delta(1-\alpha(\tau))}\sum_{t=1}^T \eta_t.\end{aligned}$$

For $\eta_t = \eta_0 d^{-1/2}t^{-3/4}$ and $\delta = \delta_0 d^{1/2}T^{-1/4}$, we have

$$\sum_{t=1}^T \varepsilon_t \leqslant \frac{1}{1-\alpha(\tau)}\left(\varepsilon_0 + \frac{4\beta_{\text{BR}}B\eta_0\sqrt{T}}{\delta_0}\right).$$

3.4.4 Proof of Theorem 3.1.3

Define $\mu_t^* = D_t(\mu_1, \theta_{\mathrm{BR}}(\mu_1), \dots, \mu_{t-1}^*, \theta_{\mathrm{BR}}(\mu_{t-1}), \xi_t)$. First we will prove that

$$\lim_{T\to\infty} \lim_{\tau\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\, \mathrm{SR}_R(\mu_t) - \mathrm{SR}_R(\mu_{\mathrm{SE}}) = 0. \tag{3.12}$$

To show this, it suffices to prove that for all t, $\mu_t \to_p \mu_t^*$ as $\tau \to \infty$. The sufficiency of this condition follows because

$$\begin{aligned}
&\lim_{T\to\infty} \lim_{\tau\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\, \mathrm{SR}_R(\mu_t) - \mathrm{SR}_R(\mu_{\mathrm{SE}}) \\
&= \lim_{T\to\infty} \lim_{\tau\to\infty} \frac{1}{T} \sum_{t=1}^{T} [\mathbb{E}\, \mathrm{SR}_R(\mu_t) - \mathbb{E}\, \mathrm{SR}_R(\mu_t^*) + \mathbb{E}\, \mathrm{SR}_R(\mu_t^*)] - \mathrm{SR}_R(\mu_{\mathrm{SE}}) \\
&= \lim_{T\to\infty} \lim_{\tau\to\infty} \frac{1}{T} \sum_{t=1}^{T} \left(\mathbb{E}\, \mathrm{SR}_R(\mu_t) - \mathbb{E}\, \mathrm{SR}_R(\mu_t^*)\right),
\end{aligned}$$

where the last step follows by the assumption that the agents play a rational strategy. Therefore, if $\mu_t \to_p \mu_t^*$, continuity of $\mathrm{SR}_R(\mu)$ implies $\mathbb{E}\, \mathrm{SR}_R(\mu_t) - \mathbb{E}\, \mathrm{SR}_R(\mu_t^*) \to 0$ and we get the desired conclusion.

We prove that $\mu_t \to_p \mu_t^*$ by induction. Notice that $\mu_1 \equiv \mu_1^*$ by definition. Suppose that $\mu_j \to_p \mu_j^*$ for all $j < t$. Denote by $\theta_{j,\tau}$ the possibly randomized algorithm that maps μ_j to θ_j. Then, for any $\mu \in \mathcal{M}$, we know that $\|\theta_{j,\tau}(\mu) - \theta_{\mathrm{BR}}(\mu)\|_2 \to_p 0$ by assumption. This in turn implies that for all $j < t$,

$$\|\theta_{j,\tau}(\mu_j) - \theta_{\mathrm{BR}}(\mu_j^*)\|_2 \leqslant \|\theta_{j,\tau}(\mu_j) - \theta_{\mathrm{BR}}(\mu_j)\|_2 + \|\theta_{\mathrm{BR}}(\mu_j) - \theta_{\mathrm{BR}}(\mu_j^*)\|_2 \to_p 0,$$

where the second term tends to zero by the continuous mapping theorem. Finally, we can apply the continuity of D_t to conclude that $\mu_t \to_p \mu_t^*$, as desired.

Let β denote the Lipschitz constant of SR_L. Finally, we we can apply this Lipschitz condition to conclude:

$$\begin{aligned}
&\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\, L(\mu_t, \theta_{t,\tau}) - L(\mu_{\mathrm{SE}}, \theta_{\mathrm{BR}}(\mu_{\mathrm{SE}})) \\
&= \frac{1}{T} \sum_{t=1}^{T} [\mathbb{E}\, L(\mu_t, \theta_{t,\tau}) \pm \mathbb{E}\, L(\mu_t, \theta_{\mathrm{BR}}(\mu_t)))] - L(\mu_{\mathrm{SE}}, \theta_{\mathrm{BR}}(\mu_{\mathrm{SE}})) \\
&= \frac{1}{T} \sum_{t=1}^{T} \left(\mathbb{E}\, L(\mu_t, \theta_{\mathrm{BR}}(\mu_t)) - L(\mu_{\mathrm{SE}}, \theta_{\mathrm{BR}}(\mu_{\mathrm{SE}})) + \mathbb{E}[L(\mu_t, \theta_{t,\tau}) - L(\mu_t, \theta_{\mathrm{BR}}(\mu_t)]\right) \\
&\leqslant \frac{\beta}{T} \sum_{t=1}^{T} \mathbb{E}\, \|\mu_t - \mu_{\mathrm{SE}}\|_2 + \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[L(\mu_t, \theta_{t,\tau}) - L(\mu_t, \theta_{\mathrm{BR}}(\mu_t)].
\end{aligned}$$

By Lemma 3.4.1, the guarantee (3.12) implies that the first term vanishes. The second term vanishes by continuity. Therefore, taking the limit over T, τ, we obtain

$$\lim_{T\to\infty}\lim_{\tau\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\,L(\mu_t,\theta_t)-L(\mu_{\mathrm{SE}},\theta_{\mathrm{BR}}(\mu_{\mathrm{SE}}))=0,$$

as desired.

3.4.5 Proof of Proposition 3.1.1

First we assume the decision-maker leads. When θ is the deployed model, the best response by the agents is to simply move by distance B in the direction of θ. Thus, $\mu_{\mathrm{BR}}(\theta)$ is given by:

$$\mu_{\mathrm{BR}}(\theta)=\arg\min_{\mu}\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}(\mu)}-x^\top\theta=\frac{\theta}{\|\theta\|_2}B.$$

This implies the following expected loss for the decision-maker:

$$\begin{aligned}L(\mu_{\mathrm{BR}}(\theta),\theta)&=\mathop{\mathbb{E}}_{z\sim\mathcal{D}\left(\frac{\theta}{\|\theta\|_2}B\right)}\ell(z;\theta)=\frac{1}{2}\mathop{\mathbb{E}}_{(x_0,y)\sim\mathcal{D}(0)}\left(y-x_0^\top\theta-\|\theta\|_2B\right)^2\\&=\frac{\sigma^2}{2}+\frac{1}{2}\|\beta-\theta\|_2^2+\frac{B^2}{2}\|\theta\|_2^2.\end{aligned}$$

This objective is convex and thus by finding a stationary point we observe that it is minimized at $\theta_{\mathrm{SE}}=\frac{\beta}{1+B^2}$. By plugging this choice back into the previous equation, we observe that the minimal Stackelberg risk of the decision-maker is equal to

$$\mathrm{SR}_L(\theta_{\mathrm{SE}})=L(\mu_{\mathrm{BR}}(\theta_{\mathrm{SE}}),\theta_{\mathrm{SE}})=\frac{\sigma^2}{2}+\frac{\|\beta\|_2^2B^2}{2(1+B^2)}. \tag{3.13}$$

Moreover, the agents' loss at θ_{SE} is equal to:

$$R(\mu_{\mathrm{BR}}(\theta_{\mathrm{SE}}),\theta_{\mathrm{SE}})=-\|\theta_{\mathrm{SE}}\|_2B=-\frac{\|\beta\|_2B}{1+B^2}.$$

Now we reverse the order of play and assume that the agents lead. If the agents move by μ, i.e. they follow the law $\mathcal{D}(\mu)$, then the decision-maker incurs loss:

$$L(\mu,\theta)=\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}(\mu)}\frac{1}{2}\left(y-x_0^\top\theta-\mu^\top\theta\right)^2=\frac{\sigma^2}{2}+\frac{1}{2}\|\beta-\theta\|_2^2+\frac{1}{2}(\mu^\top\theta)^2.$$

By computing a stationary point, we find that the best response of the decision-maker is:

$$\theta_{\mathrm{BR}}(\mu)=(I+\mu\mu^\top)^{-1}\beta=\left(I-\frac{\mu\mu^\top}{1+\|\mu\|_2^2}\right)\beta.$$

The Stackelberg risk of the strategic agent is then

$$\begin{aligned}\mathrm{SR}_R(\mu) = R(\mu, \theta_{\mathrm{BR}}(\mu)) = -\mu^\top \theta_{\mathrm{BR}}(\mu) = -\mu^\top \left(I - \frac{\mu\mu^\top}{1+\|\mu\|_2^2}\right)\beta \\ = -\mu^\top\beta + \frac{\|\mu\|_2^2 \mu^\top \beta}{1+\|\mu\|_2^2} = -\frac{\mu^\top\beta}{1+\|\mu\|_2^2}.\end{aligned}$$

Among all μ such that $\|\mu\|_2 = C$, $\mathrm{SR}_R(\mu)$ is minimized when μ points in the β direction: $\mu = C\frac{\beta}{\|\beta\|_2}$. With this reparameterization, we can equivalently write $\min_\mu \mathrm{SR}_R(\mu)$ as

$$\min_{C>0} \|\beta\|_2 \frac{-C}{1+C^2}.$$

This function is decreasing for $C \in (0,1]$, and increasing for $C > 1$. Therefore, $\mu_{\mathrm{SE}} = \min(1,B)\frac{\beta}{\|\beta\|_2}$, and

$$\mathrm{SR}_R(\mu_{\mathrm{SE}}) = -\|\beta\|_2 \frac{\min(1,B)}{1+\min(1,B)^2}.$$

Finally, we evaluate the decision-maker's loss at μ_{SE}:

$$\begin{aligned}L(\mu_{\mathrm{SE}}, \theta_{\mathrm{BR}}(\mu_{\mathrm{SE}})) &= \frac{\sigma^2}{2} + \frac{1}{2}\frac{(\beta^\top\mu_{\mathrm{SE}})^2\|\mu_{\mathrm{SE}}\|_2^2}{(1+\|\mu_{\mathrm{SE}}\|_2^2)^2} + \frac{1}{2}\left(\|\beta\|_2\frac{\min(1,B)}{1+\min(1,B)^2}\right)^2 \\ &= \frac{\sigma^2}{2} + \frac{1}{2}\frac{\|\beta\|_2^2\min(1,B)^4}{(1+\min(1,B)^2)^2} + \frac{1}{2}\left(\|\beta\|_2\frac{\min(1,B)}{1+\min(1,B)^2}\right)^2 \\ &= \frac{\sigma^2}{2} + \frac{\|\beta\|_2^2\min(1,B)^2}{2(1+\min(1,B)^2)}.\end{aligned}$$

3.4.6 Proof of Proposition 3.1.2

First we evaluate $L(\mu,\theta)$:

$$\begin{aligned}L(\mu,\theta) &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}(\mu)}\left[-yx^\top\theta + \log(1+e^{x^\top\theta})\right] \\ &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}(\mu)}\left[\log(e^{-yx^\top\theta} + e^{(1-y)x^\top\theta})\right] \\ &= \mathop{\mathbb{E}}_{(x_0,y)\sim\mathcal{D}(0)}[\mathbf{1}\{y=1\}\log(1+e^{-x_0^\top\theta}) + \mathbf{1}\{y=0\}\log(1+e^{x_0^\top\theta+\mu^\top\theta})].\end{aligned}$$

We prove that the agents are never worse off if they lead. We will provide a sufficient condition; namely, we will show that

$$\mathrm{SR}_R\left(\frac{\theta_{\mathrm{SE}}}{\|\theta_{\mathrm{SE}}\|_2}B\right) = R(\mu_{\mathrm{BR}}(\theta_{\mathrm{SE}}), \theta_{\mathrm{SE}}).$$

This immediately implies that $\mathrm{SR}_R(\mu_{\mathrm{SE}}) \leqslant R(\mu_{\mathrm{BR}}(\theta_{\mathrm{SE}}), \theta_{\mathrm{SE}})$.

To see this, first observe that

$$R(\mu_{\mathrm{BR}}(\theta_{\mathrm{SE}}), \theta_{\mathrm{SE}}) = B\|\theta_{\mathrm{SE}}\|_2,$$

where $\theta_{\mathrm{SE}} = \arg\min_\theta L\left(\frac{\theta}{\|\theta\|_2}B, \theta\right)$. Here we use the fact that the best response of the agents is to simply move by distance B in the direction of θ:

$$\mu_{\mathrm{BR}}(\theta) = \arg\max_{\mu\in\mathcal{M}} \theta^\top \mu = \frac{\theta}{\|\theta\|_2}B.$$

By the fact that θ_{SE} is a Stackelberg equilibrium, we know that $\nabla_\theta \mathrm{SR}_L(\theta_{\mathrm{SE}}) = 0$, where $\mathrm{SR}_L(\theta) = L(\frac{\theta}{\|\theta\|_2}B, \theta)$:

$$\nabla_\theta \mathrm{SR}_L(\theta) = \mathop{\mathbb{E}}_{(x_0,y)\sim\mathcal{D}(0)}\left[\mathbf{1}\{y=1\}\frac{e^{-x_0^\top\theta}(-x_0)}{1+e^{-x_0^\top\theta}} + \mathbf{1}\{y=0\}\frac{e^{x_0^\top\theta+\|\theta\|_2 B}(x_0+\frac{\theta}{\|\theta\|_2}B)}{1+e^{x_0^\top\theta+\|\theta\|_2 B}}\right].$$

In contrast, consider $\nabla_\theta L(\mu, \theta)$:

$$\nabla_\theta L(\mu,\theta) = \mathop{\mathbb{E}}_{(x_0,y)\sim\mathcal{D}(0)}\left[\mathbf{1}\{y=1\}\frac{e^{-x_0^\top\theta}(-x_0)}{1+e^{-x_0^\top\theta}} + \mathbf{1}\{y=0\}\frac{e^{x_0^\top\theta+\mu^\top\theta}(x_0+\mu)}{1+e^{x_0^\top\theta+\mu^\top\theta}}\right].$$

Notice that $\nabla_\theta \mathrm{SR}_L(\theta_{\mathrm{SE}}) = 0$ implies that $\nabla_\theta L(\frac{\theta_{\mathrm{SE}}}{\|\theta_{\mathrm{SE}}\|_2}B, \theta_{\mathrm{SE}}) = 0$. Since $L(\mu,\theta)$ is convex in θ, this condition implies that θ_{SE} is a best response to $\frac{\theta_{\mathrm{SE}}}{\|\theta_{\mathrm{SE}}\|_2}B$, hence

$$\mathrm{SR}_R\left(\frac{\theta_{\mathrm{SE}}}{\|\theta_{\mathrm{SE}}\|_2}B\right) = R(\mu_{\mathrm{BR}}(\theta_{\mathrm{SE}}), \theta_{\mathrm{SE}}).$$

Now we analyze the decision-maker's preference. Notice that $L(\mu,\theta)$ is increasing in $\mu^\top\theta$; that is, for any θ it holds that $\max_\mu L(\mu,\theta) = L(\mu_{\mathrm{BR}}(\theta), \theta)$. Using this, we observe that for every θ we have

$$L(\mu_{\mathrm{SE}}, \theta_{\mathrm{BR}}(\mu_{\mathrm{SE}})) \leqslant L(\mu_{\mathrm{SE}}, \theta) \leqslant \max_{\mu\in\mathcal{M}} L(\mu,\theta) = \mathrm{SR}_L(\theta).$$

Since this also holds for $\theta = \theta_{\mathrm{SE}}$, we conclude that following is never worse than leading for the decision-maker.

3.4.7 Proof of Theorem 3.2.1

First consider the case $\varepsilon = 0$. We start with a sufficient condition for a target classification outcome. For a point $x \in \mathcal{X}$, we define

$$\Delta_x = \max_{y\in\mathcal{Y}} \mathcal{D}_0(y|x) - \mathcal{D}_0(y^*|x)$$

as the suboptimality of a target class on the base data.

Claim 3.4.1. *For any $x \in \mathcal{X}$, we have $f(x) = y^*$ provided that $\alpha > (1-\alpha)\Delta_x \mathcal{D}_0(x)/\mathcal{D}^*(x)$.*

Proof. Note that $f(x) = y^*$ if, for every $y \neq y^*$, $\mathcal{D}(y^*|x) > \mathcal{D}(y|x)$. Equivalently, $\mathcal{D}(x, y^*) - \mathcal{D}(x, y) > 0$. But,

$$\mathcal{D}(x, y^*) = \alpha\mathcal{D}^*(x, y^*) + (1-\alpha)\mathcal{D}_0(x, y^*) = \alpha\mathcal{D}^*(x) + (1-\alpha)\mathcal{D}_0(y^*|x)\mathcal{D}_0(x)$$

In the last step we used the fact that all labels in the support of $\mathcal{D}^*$ equal y^*. Similarly, for $y \neq y^*$,

$$\mathcal{D}(x, y) = \alpha\mathcal{D}^*(x, y) + (1-\alpha)\mathcal{D}_0(x, y) = (1-\alpha)\mathcal{D}_0(y|x)\mathcal{D}_0(x)\,.$$

The claim follows by rearranging terms and dividing both sides by $\mathcal{D}^*(x)$. □

Now,

$$\begin{aligned} S(\alpha) &= \mathbb{P}_{x\sim\mathcal{D}_0}\left\{f(g(x)) = y^*\right\} \\ &= \mathbb{P}_{x\sim\mathcal{D}^*}\left\{f(x) = y^*\right\} \\ &\geqslant \mathbb{P}_{x\sim\mathcal{D}^*}\left\{\alpha > (1-\alpha)\frac{\mathcal{D}_0(x)}{\mathcal{D}^*(x)}\Delta_x\right\} && \text{(Claim 3.4.1)} \\ &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}^*} \mathbf{1}\left\{1 - \frac{(1-\alpha)}{\alpha}\frac{\mathcal{D}_0(x)}{\mathcal{D}^*(x)}\Delta_x > 0\right\} \\ &\geqslant \mathop{\mathbb{E}}_{x\sim\mathcal{D}^*}\left[1 - \frac{(1-\alpha)}{\alpha}\frac{\mathcal{D}_0(x)}{\mathcal{D}^*(x)}\Delta_x\right] \\ &= 1 - \frac{1-\alpha}{\alpha}\mathop{\mathbb{E}}_{x\sim\mathcal{D}^*}\left[\frac{\mathcal{D}_0(x)}{\mathcal{D}^*(x)}\Delta_x\right] \\ &\geqslant 1 - \frac{1-\alpha}{\alpha}\mathcal{D}_0(\mathcal{X}^*)\Delta\,, \end{aligned}$$

where the last step uses the definition $\Delta = \max_{x\in\mathcal{X}^*}\Delta_x$.

Consider $\varepsilon > 0$**.** By Lemma 3.4.2, we have that $\mathcal{D}'(x, y^*) > \mathcal{D}'(x, y)$, meaning $f(x) = y^*$, provided that $\mathcal{D}(x, y^*) > \mathcal{D}(x, y) + \frac{\varepsilon}{1-\varepsilon}$. Repeating the steps in the proof for $\varepsilon = 0$ with the additional $\varepsilon/(1-\varepsilon)$ term, we conclude that

$$S(\alpha) \geqslant 1 - \frac{\varepsilon}{1-\varepsilon} - \frac{1-\alpha}{\alpha}\mathcal{D}_0(\mathcal{X}^*)\Delta.$$

3.4.8 Proof of Theorem 3.2.2

We prove the case where $\varepsilon = 0$. The extension to $\varepsilon > 0$ follows as in Theorem 3.2.1.

Claim 3.4.2. *Fix a point $x^* \in \mathcal{X}^*$ in the signal set. We have $f(x^*) = y^*$ provided that*

$$\alpha > \frac{1-p}{p}\frac{\mathcal{D}_0(x^*)}{\mathcal{D}_0(g^{-1}(x^*))}.$$

Here, $g^{-1}(x^*) = \{x \in \mathcal{X} : g(x) = x^*\}$.

Proof. For $f(x^*) = y^*$ to hold, we need $\mathcal{D}(y^*|x^*) > \max_{y \neq y^*} \mathcal{D}(y|x^*)$. Equivalently, $\mathcal{D}(x^*, y^*) > \max_{y \neq y^*} \mathcal{D}(x^*, y)$.

By the definition of the feature-only signal strategy and the assumption that $\mathcal{D}_0(y^*|x) \geqslant p$ for all $x \in \mathcal{X}$, each point $x \in g^{-1}(x^*)$ must have $\mathcal{D}_0(y^*|x) \geqslant p$. Hence, for all $x^* \in \mathcal{X}^*$,

$$\mathcal{D}(x^*, y^*) = \alpha \mathcal{D}^*(x^*, y^*) + (1 - \alpha)\mathcal{D}_0(x^*, y^*) \geqslant \alpha p \mathcal{D}_0(g^{-1}(x^*))\,.$$

On the other hand, for every $y \neq y^*$, we must have

$$\mathcal{D}(x^*, y) = \mathcal{D}_0(x^*, y) = \mathcal{D}_0(y|x^*)\mathcal{D}_0(x^*) \leqslant (1 - p)\mathcal{D}_0(x^*)\,.$$

The claim follows by rearranging. □

We can lower bound the success rate as

$$\begin{aligned} S(\alpha) &= \mathbb{P}_{x \sim \mathcal{D}_0} \{f(g(x)) = y^*\} \\ &= \sum_{x^* \in \mathcal{X}^*} \mathbb{P}_{x \sim \mathcal{D}_0} \left\{f(g(x)) = y^* \mid x \in g^{-1}(x^*)\right\} \mathbb{P}_{x \sim \mathcal{D}_0}\{x \in g^{-1}(x^*)\} \\ &= \sum_{x^* \in \mathcal{X}^*} \mathbf{1}\{f(x^*) = y^*\}\, \mathcal{D}_0(g^{-1}(x^*))\,. \end{aligned} \tag{3.14}$$

Proceeding for fixed $x^* \in \mathcal{X}^*$,

$$\begin{aligned} \mathbf{1}\{f(x^*) = y^*\} &\geqslant \mathbf{1}\left\{\alpha > \frac{1-p}{p} \frac{\mathcal{D}_0(x^*)}{\mathcal{D}_0(g^{-1}(x^*))}\right\} && \text{(Claim 3.4.2)} \\ &= \mathbf{1}\left\{1 - \frac{1-p}{p\alpha} \frac{\mathcal{D}_0(x^*)}{\mathcal{D}_0(g^{-1}(x^*))} > 0\right\} \\ &\geqslant 1 - \frac{1-p}{p\alpha} \cdot \frac{\mathcal{D}_0(x^*)}{\mathcal{D}_0(g^{-1}(x^*))}. \end{aligned}$$

Plugging this back into (3.14),

$$\begin{aligned} \mathbb{P}_{x \sim \mathcal{D}_0} \{f(g(x)) = y^*\} &= 1 - \frac{1-p}{p\alpha} \sum_{x^* \in \mathcal{X}^*} \frac{\mathcal{D}_0(x^*)}{\mathcal{D}_0(g^{-1}(x^*))} \cdot \mathcal{D}_0(g^{-1}(x^*)) \\ &\geqslant 1 - \frac{1-p}{p\alpha} \mathcal{D}_0(\mathcal{X}^*)\,. \end{aligned}$$

3.4.9 Proof of Theorem 3.2.3

We again prove the case where $\varepsilon = 0$. The extension to $\varepsilon > 0$ follows by invoking Lemma 3.4.2, as in Theorem 3.2.1.

We start from the following claim.

Claim 3.4.3. *For any $x \in \mathcal{X}$ we have $f(x) = f(g(x))$ provided that*

$$\alpha > (1-\alpha)2\tau(x),$$

where $\tau(x) = \max_{y\in\mathcal{Y}} |\mathcal{D}_0(y|x) - \mathcal{D}_0(y|g(x))|$.

Proof. Denote $y^*(x) = \arg\max_{y\in\mathcal{Y}} \mathcal{D}_0(y|g(x))$. By construction of the strategy we know that $f(g(x)) = y^*(x)$ and it remains to prove that $f(x) = y^*(x)$ under the condition of the claim.

We have $f(x) = y^*(x)$ if $\mathcal{D}(y^*(x)|x) > \mathcal{D}(y|x)$ for any $y \neq y^*(x)$. We have

$$\begin{aligned}\mathcal{D}(y^*(x)|x) &= (1-\alpha)\mathcal{D}_0(y^*(x)|x) + \alpha\mathcal{D}^*(y^*(x)|x) = (1-\alpha)\mathcal{D}_0(y^*(x)|x) + \alpha,\\ \mathcal{D}(y|x) &= (1-\alpha)\mathcal{D}_0(y|x) + \alpha\mathcal{D}^*(y|x) = (1-\alpha)\mathcal{D}_0(y|x),\end{aligned}$$

where we used that the erasure strategy implies $\mathcal{D}^*(y^*(x)|x) = 1$. Together this means that, when

$$\alpha > (1-\alpha)\left[\max_{y\in\mathcal{Y}} \mathcal{D}_0(y|x) - \mathcal{D}_0(y^*(x)|x)\right],$$

then $f(x) = y^*(x)$. Using the definition of $y^*(x)$, we can bound the right-hand side by

$$\begin{aligned}\mathcal{D}_0(y|x) - \mathcal{D}_0(y^*(x)|x) &\leqslant \mathcal{D}_0(y|x) - \mathcal{D}_0(y|g(x)) + \mathcal{D}_0(y^*(x)|g(x)) - \mathcal{D}_0(y^*(x)|x)\\ &\leqslant 2\tau(x).\end{aligned}$$

The claim follows. □

It remains to bound the success of the strategy:

$$\begin{aligned}\mathrm{SR}(\alpha) &= \mathbb{P}_{x\sim\mathcal{D}_0}\{f(x) = f(g(x))\}.\\ &= \mathbb{P}_{x\sim\mathcal{D}_0}\{f(x) = y^*(x)\}.\\ &\geqslant \mathbb{P}_{x\sim\mathcal{D}_0}\{\alpha > (1-\alpha)2\tau(x)\}\\ &= \mathbb{P}_{x\sim\mathcal{D}_0}\left\{1 - \frac{1-\alpha}{\alpha}2\tau(x) > 0\right\}\\ &\geqslant \mathop{\mathbb{E}}_{x\sim\mathcal{D}_0}\left[1 - \frac{2(1-\alpha)}{\alpha}\cdot\tau(x)\right]\\ &= 1 - \frac{2(1-\alpha)}{\alpha}\cdot\tau,\end{aligned}$$

where we use the fact that $\tau = \mathbb{E}_{x\sim\mathcal{D}_0}\,\tau(x)$.

3.4.10 Proof of Theorem 3.2.4

Let $\mathcal{D}'$ be a gradient-cancelling distribution for θ^*. Denote $p = \min\left(1, \frac{1}{\alpha}\frac{\|g_{\mathcal{D}_0}(\theta^*)\|}{\|g_{\mathcal{D}'}(\theta^*)\|+\|g_{\mathcal{D}_0}(\theta^*)\|}\right)$. Then,

$$\begin{aligned}
\mathop{\mathbb{E}}_{z\sim\mathcal{D}} \nabla\ell(\theta^*; z) &= (1-\alpha)\mathop{\mathbb{E}}_{z\sim\mathcal{D}_0} \nabla\ell(\theta^*; z) + \alpha \mathop{\mathbb{E}}_{z\sim\mathcal{D}^*} \nabla\ell(\theta^*; z)\\
&= (1-\alpha p)\mathop{\mathbb{E}}_{z\sim\mathcal{D}_0} \nabla\ell(\theta^*; z) + \alpha p \mathop{\mathbb{E}}_{z\sim\mathcal{D}'} \nabla\ell(\theta^*; z)\\
&= (1-\alpha p) g_{\mathcal{D}_0}(\theta^*) + \alpha p\ g_{\mathcal{D}'}(\theta^*)\\
&= \left(1-\alpha p - \alpha p\frac{\|g_{\mathcal{D}'}(\theta^*)\|}{\|g_{\mathcal{D}_0}(\theta^*)\|}\right) g_{\mathcal{D}_0}(\theta^*)\\
&= \left(1-\alpha p\frac{\|g_{\mathcal{D}_0}(\theta^*)\| + \|g_{\mathcal{D}'}(\theta^*)\|}{\|g_{\mathcal{D}_0}(\theta^*)\|}\right) g_{\mathcal{D}_0}(\theta^*)\\
&= \max\left(1-\alpha\frac{\|g_{\mathcal{D}_0}(\theta^*)\| + \|g_{\mathcal{D}'}(\theta^*)\|}{\|g_{\mathcal{D}_0}(\theta^*)\|}, 0\right) g_{\mathcal{D}_0}(\theta^*)\\
&= \max\left((1-\alpha)\|g_{\mathcal{D}_0}(\theta^*)\| - \alpha\|g_{\mathcal{D}'}(\theta^*)\|, 0\right) \frac{g_{\mathcal{D}_0}(\theta^*)}{\|g_{\mathcal{D}_0}(\theta^*)\|}.
\end{aligned}$$

Therefore, $\|\mathbb{E}_{z\sim\mathcal{D}} \nabla\ell(\theta^*; z)\| = \max\left((1-\alpha)\|g_{\mathcal{D}_0}(\theta^*)\| - \alpha\|g_{\mathcal{D}'}(\theta^*)\|, 0\right)$. Applying the definition of μ-strong convexity, we get

$$\begin{aligned}
\|\theta^* - \theta\| &\leqslant \frac{1}{\mu}\|\mathop{\mathbb{E}}_{z\sim\mathcal{D}} \nabla\ell(\theta^*; z) - \mathop{\mathbb{E}}_{z\sim\mathcal{D}} \nabla\ell(\theta; z)\|\\
&= \frac{1}{\mu}\|\mathop{\mathbb{E}}_{z\sim\mathcal{D}} \nabla\ell(\theta^*; z)\|\\
&= \frac{1}{\mu}\max\left((1-\alpha)\|g_{\mathcal{D}_0}(\theta^*)\| - \alpha\|g_{\mathcal{D}'}(\theta^*)\|, 0\right).
\end{aligned}$$

The first equality follows because $\mathbb{E}_{z\sim\mathcal{D}} \nabla\ell(\theta; z) = 0$ due to the loss being convex and the firm being a risk minimizer. Multiplying both sides by -1, we obtain a lower bound on the success $S(\alpha) = -\|\theta^* - \theta\|$.

3.4.11 Proof of Corollary 3.2.4

To achieve $S(\alpha) = 0$, Theorem 3.2.4 shows that it suffices to have $\alpha\|g_{\mathcal{D}'}(\theta^*)\| = (1-\alpha)\|g_{\mathcal{D}_0}(\theta^*)\|$, for any μ. Rearranging the terms and expressing α completes the proof.

3.4.12 Proof of Proposition 3.2.1

If u is convex, then for all θ' we know

$$u(\theta') \geqslant u(\theta_0) + \nabla u(\theta_0)^\top(\theta' - \theta_0).$$

Let $\theta^* = \theta_0 + \frac{\nabla u(\theta_0)}{\|\nabla u(\theta_0)\|^2} U$. Then, $u(\theta^*) - u(\theta_0) \geqslant U$.

Now, we apply Corollary 3.2.4 to upper bound the critical mass needed to reach θ^*. We have

$$\alpha^* \leqslant \frac{\|g_{\mathcal{D}_0}(\theta^*)\|}{\|g_{\mathcal{D}'}(\theta^*)\| + \|g_{\mathcal{D}_0}(\theta^*)\|} \leqslant \frac{\beta\|\theta^* - \theta_0\|}{g_{\text{lb}} + \beta\|\theta^* - \theta_0\|},$$

where we apply the fact that the loss is smooth and the definition of g_{lb}. Observing that $\|\theta^* - \theta_0\| = \frac{U}{\|\nabla u(\theta_0)\|}$ completes the proof.

3.4.13 Proof of Theorem 3.2.5

Fix a time step t and a model θ_t. Denote by $\mathcal{D}'_t$ the gradient-redirecting distribution found at step t and let $\xi(\theta_t) = \frac{\|g_{\mathcal{D}'_t}(\theta_t) + \frac{1-\alpha}{\alpha} g_{\mathcal{D}_0}(\theta_t)\|}{\|\theta_t - \theta^*\|}$. Then, the gradient-redirecting strategy induces the following gradient evaluated on $\mathcal{D}_t$:

$$\begin{aligned} g_{\mathcal{D}_t}(\theta_t) &= \alpha g_{\mathcal{D}'_t}(\theta_t) + (1-\alpha) g_{\mathcal{D}_0}(\theta_t) \\ &= -\alpha \frac{1-\alpha}{\alpha} g_{\mathcal{D}_0}(\theta_t) + \alpha \xi(\theta_t)(\theta_t - \theta^*) + (1-\alpha) g_{\mathcal{D}_0}(\theta_t) \\ &= \alpha \xi(\theta_t)(\theta_t - \theta^*). \end{aligned}$$

Now let $c = \min_{\lambda \in [0,1]} \xi(\lambda\theta_0 + (1-\lambda)\theta^*)$. Applying the strategy repeatedly across time steps yields

$$\begin{aligned} \|\theta_T - \theta^*\| &\leqslant \|\theta_{T-1} - \eta\alpha\xi(\theta_{T-1})(\theta_{T-1} - \theta^*) - \theta^*\| \\ &\leqslant (1 - \eta\alpha\xi(\theta_{T-1}))\|\theta_{T-1} - \theta^*\| \\ &\leqslant (1 - \eta\alpha c)\|\theta_{T-1} - \theta^*\| \\ &\leqslant (1 - \eta\alpha c)^T \|\theta_0 - \theta^*\|, \end{aligned}$$

hence $S_T(\alpha) = -\|\theta_T - \theta^*\| \geqslant -(1-\eta\alpha c)^T\|\theta_0 - \theta^*\|$. Setting $C(\alpha) = \alpha c$ concludes the proof.

Part II

Statistical Inference in Feedback Loops

Chapter 4

Selective Inference

Modern scientific investigations increasingly involve choosing inferential questions of interest only *after* seeing the data. While this practice offers more freedom to the scientist than the traditional paradigm of specifying the relevant questions up front, it is by now well understood that it also creates undesirable *selection bias*, thereby invalidating type I error guarantees of classical statistical methods. The area of *selective inference* formally describes this problem and offers rigorous solutions across a variety of settings.

One standard solution is to perform *simultaneous inference*, i.e., deliver valid answers to all questions that could possibly be asked. However, simultaneous inference can be unnecessarily conservative when many questions, though possible, are unlikely to be of interest in the first place. For example, suppose that a clinical trial estimates the effectiveness of multiple treatments and, after observing the data, it is clear that there are many ineffective treatments and only a handful of effective ones. Even if we are only interested in constructing a confidence interval for the effectiveness of the seemingly best treatment, simultaneous inference would still account for the possible estimation error for the clearly ineffective treatments.

Another solution is to perform *conditional* selective inference, which delivers valid answers after conditioning on the event that a specific selection was made. Unlike simultaneous inference, conditional selective inference adapts to the specifics of the selection criterion and how “obvious” the selection is; however, implementing a conditional correction generally requires a tractable characterization of the selection method and parametric assumptions, constraining the applicability of the approach. Moreover, the conditional nature of the inferences can be overly conservative, even more conservative than simultaneous inference.

In this chapter we introduce two broadly applicable principles for providing valid selective inferences that are neither simultaneous nor conditional. One is based on quantifying the *algorithmic stability* of the selection and the other is based on a *locally simultaneous correction*, that is, a correction only over selections that seem plausible in hindsight. Both approaches come with rigorous error guarantees and allow for powerful, selection-adaptive inferences. Moreover, they are nonparametrically applicable and computationally tractable.

The material in this chapter is based on works co-authored with William Fithian and Michael I. Jordan [192, 193].

4.1 Preliminaries

To build intuition, we begin by presenting two motivating examples of selective inference problems. Then, we introduce the general formal setup we will be working within.

4.1.1 Motivating vignettes

Vignette 1: Winner's curse. The first vignette considers the problem of selecting the largest observed effect to do inference on. Suppose that we observe an m-dimensional vector $y \sim \mathcal{N}(\mu, \sigma^2 I)$. The m coordinates of y can correspond to, for example, the effectiveness of m different treatments, in which case the selection corresponds to focusing on the seemingly best treatment. The m outcomes can also correspond to measurements of a time series over m time steps, in which case selection focuses on the time step at which the series achieves extreme values. Finally, y can capture an estimate of the effectiveness of a treatment on m different subgroups (e.g., m age groups); the selection would then ask for the effectiveness within the subgroup for which the treatment seems most promising.

We are interested in doing inference on the most significant effect; formally, denoting $\hat{\gamma} = \arg\max_{\gamma \in [m]} y_\gamma$, we want to construct a confidence interval for $\hat{\gamma}$. Note that this is a *random* inferential target because $\hat{\gamma}$ is a function of the data.

One simple way of providing valid inference for $\mu_{\hat{\gamma}}$ is to apply the Bonferroni correction:

$$P_\mu\{\mu_{\hat{\gamma}} \in (y_{\hat{\gamma}} \pm z_{1-\alpha/(2m)}\sigma)\} \geqslant 1 - \alpha,$$

where z_q is the q quantile of the standard normal distribution.

Benjamini et al. [9] show that a tighter correction is valid, namely

$$P_\mu\{\mu_{\hat{\gamma}} \in (y_{\hat{\gamma}} \pm z_{1-\alpha/(m+1)}\sigma)\} \geqslant 1 - \alpha.$$

While the Benjamini et al. correction is tighter, neither strategy is *data-adaptive*; that is, the stated confidence interval widths do not depend on how "obvious" the winner is. Intuitively, if the winner stands out, then there is little true selection: even if we obtained an independent sample of the data, the winner would probably stay the same. This in turn means that, if the winner is obvious, the inferential target is *essentially fixed* a priori and we should expect the confidence intervals to approach nominal, uncorrected intervals: $(y_{\hat{\gamma}} \pm z_{1-\alpha/2}\sigma)$.

In this chapter we will propose two strategies that adapt to the data at hand. When the winner stands out, both strategies will be able to return nearly uncorrected confidence intervals for the winning effect. More generally, the two strategies will only correct for "plausible" winners, as we will make more concrete later. Moreover, we note that our solutions will be applicable even when the errors are not Gaussian; they will be applicable even nonparametrically.

Vignette 2: Feature selection. In the second example we look at inference after data-driven feature selection. In virtually all domains of statistical applications, feature selection is widely taught and practiced, and even stands as a research area of its own. Sometimes feature selection is even unavoidable; in the canonical setting of linear regression, the statistician often starts with a pool of candidate variables large enough that it makes the solution unidentifiable without additional constraints.

To describe the problem formally, suppose we have a fixed design matrix, $X \in \mathbb{R}^{n\times d}$, with n observations and d features and a corresponding outcome vector $y \sim \mathcal{N}(\mu, \sigma^2 I) \in \mathbb{R}^n$. Denote by X_i the columns of X, for $i \in [d]$. We want to select a *model* $\hat{M} \equiv \hat{M}(y)$ corresponding to a subset of the d features, and then regress the outcome onto the selected features. Following the proposal of Berk et al. [12], for a fixed model $M \subseteq [d]$ the so-called *projection parameter* is the target of inference. This parameter is obtained by approximating the outcome using the columns in X indexed by M:

$$\theta_M := \arg\min_{\theta} \mathbb{E}\, \|y - X_M\theta\|_2^2 = X_M^+\mu,$$

where X_M^+ denotes the pseudoinverse of X_M. We denote the empirical counterpart of θ_M by $\hat{\theta}_M := X_M^+ y$. For a data-driven choice of model $\hat{M}$, the inferential target is therefore $\theta_{\hat{M}}$. We use $\theta_{j\cdot M}$ to denote the entry of θ_M corresponding to feature j. Note that, in general, $\theta_{j\cdot M} \neq \theta_{j\cdot M'}$ for two different models M, M'.

Therefore, the goal is to construct intervals C_i such that

$$P_\mu\{\theta_{i\cdot\hat{M}} \in C_i, \forall i \in \hat{M}\} \geqslant 1-\alpha.$$

Berk et al. [12] provide one solution to this problem, called the *PoSI correction*, which relies on taking a *simultaneous* correction over all possible estimands we could ever ask about. Mathematically, they compute a width parameter q_{PoSI} such that

$$P_\mu\{\theta_{i\cdot M} \in (\hat{\theta}_{i\cdot M} \pm q_{\mathrm{PoSI}} \cdot \hat{\sigma}_{i\cdot M}), \forall i \in M, \forall M \in \mathcal{M}\} \geqslant 1-\alpha,$$

where $\hat{\sigma}_{i\cdot M} = \sigma\sqrt{((X_M^\top X_M)^{-1})_{ii}}$ is the usual standard error term in linear regression and $\mathcal{M}$ is the space of *all possible models* (often all $2^{[d]}$ subsets of the features). Again, we see that this correction is not data-adaptive. Indeed, since it asks for a correction over a large number of models, usually it is overly conservative. For the same reason, computing q_{PoSI} is computationally challenging, as it requires searching over all possible models.

An alternative solution to inference after feature selection is data splitting: we use a fraction $f \in (0,1)$ of the data for selection and the remaining $1-f$ fraction for inference. Data splitting is appealing because, if the two subsets of the data are independent, classical inferences will be valid regardless of the selection procedure. However, data splitting is not universally applicable as one cannot always obtain two independent data sets, and even if applicable, it can suffer a significant loss in power, such as when only a few samples capture some relevant information.

The solutions we will provide in this chapter will address the problem of inference after feature selection in a computationally efficient manner, and they will often be more powerful than the two described baselines. Finally, we emphasize that our solutions will be applicable even when data splitting is not an option, such as when there are spatial or temporal dependencies in the data.

4.1.2 Formal setup

We consider a possibly nonparametric family of distributions $\mathcal{P}$. For every distribution $P \in \mathcal{P}$, we have a family of possible target estimands indexed by $\gamma \in \Gamma$, $\{\theta_\gamma(P)\}_{\gamma\in\Gamma}$. For example, $\mathcal{P} = \{\mathcal{N}(\mu, I_m) : \mu \in \mathbb{R}^m\}$ could be a location family, $\Gamma = \{1, \dots, m\}$ the set of possible target indices, and $\theta_\gamma(\mathcal{N}(\mu, I_m)) = \mu_\gamma$ asks for the coordinate of μ indexed by γ. The relevant distribution P will usually be clear from the context, in which case we will simplify notation and write $\theta_\gamma \equiv \theta_\gamma(P)$.

Selective inference studies the problem of doing inference on $\{\theta_\gamma : \gamma \in \widehat{\Gamma}(y)\}$ given data $y \sim P$, where $\widehat{\Gamma}(y)$ determines a *data-dependent* set of inferential targets. We will adopt the convention that $\widehat{\Gamma} \equiv \widehat{\Gamma}(y)$ when the argument y is clear from the context. When there is a single selected target, we will denote it by $\hat{\gamma} \equiv \hat{\gamma}(y)$; in that case $\widehat{\Gamma} = \{\hat{\gamma}\}$. The goal is to construct confidence intervals for the selected targets, $\{C_\gamma\}_{\gamma\in\widehat{\Gamma}}$, such that

$$P\{\theta_\gamma \in C_\gamma, \forall \gamma \in \widehat{\Gamma}\} \geqslant 1 - \alpha,$$

where $\alpha \in (0, 1)$ is a pre-specified error level.

4.2 Existing solutions

Most existing solutions to the problem of selective inference fall under one of two categories: *simultaneous* approaches and *conditional* approaches.

Simultaneous approaches. The basic principle of simultaneous approaches is to ensure valid inferences for *all* questions that could possibly be asked. More formally, if we denote by C_γ a confidence region for target $\gamma \in \Gamma$, then the basic principle of simultaneous inference is captured by the inequality:

$$P\{\theta_{\hat{\gamma}} \notin C_{\hat{\gamma}}\} \leqslant P\{\exists \gamma \in \Gamma : \theta_\gamma \notin C_\gamma\}.$$

Simultaneous approaches construct C_γ so that the right-hand side is bounded at a pre-specified level $\alpha \in (0, 1)$. Notice that the right-hand side has no dependence on $\hat{\gamma}$. Indeed, simultaneous approaches ensure valid selective inference in a selection-agnostic manner and as such are broadly applicable. Canonical examples of simultaneous inference methods include the Bonferroni correction, Holm's procedure [80], and other related extensions [79, 81].

In the context of multivariate normal observations, simultaneous inference typically relies on estimating quantiles of the maximal z- or t-statistic [12, 19, 65, 66, 82]. See [47] for an overview of simultaneous inference methods.

We give examples of two simultaneous inference methods that are common choices for nonparametric problems and parametric problems, respectively.

Example 4.2.1 (Bonferroni correction)**.** *The Bonferroni correction achieves simultaneous control by using nominal (i.e., unadjusted) intervals at the corrected error level $\alpha/|\Gamma|$:*

$$C_\gamma^{Bonf(\alpha)} = C_\gamma^{nom(\alpha/|\Gamma|)}.$$

Here, $C_\gamma^{nom(\alpha)}$ are any intervals that satisfy

$$P\left\{\theta_\gamma \in C_\gamma^{nom(\alpha)}\right\} \geqslant 1-\alpha,$$

for a specified level $\alpha \in (0,1)$. Bonferroni-corrected intervals can be applied nonparametrically and are valid regardless of any dependencies between the different estimation problems included in Γ.

Example 4.2.2 (Maximal z- or t-statistic)**.** *If we have prior knowledge about the dependence structure of the different estimation problems included in Γ, there are approaches that outperform the Bonferroni correction. Suppose that for each $\gamma \in \Gamma$ we observe $\hat\theta_\gamma \sim \mathcal{N}(\theta_\gamma, \sigma_\gamma^2)$ and that jointly these observations make a multivariate Gaussian vector with a known covariance matrix. Denote the known covariance matrix of $(\hat\theta_\gamma)_{\gamma\in\Gamma}$ by Σ. Then, standard simultaneous confidence intervals are obtained by simulating the $1-\alpha$ quantile of the maximal z-statistic given by:*

$$\max_{\gamma\in\Gamma} \frac{|Z_\gamma|}{\sigma_\gamma},$$

where $(Z_\gamma)_{\gamma\in\Gamma} \sim \mathcal{N}(0,\Sigma)$. Denote this quantile by q. We construct the confidence intervals as

$$C_\gamma^\alpha = \left(\hat\theta_\gamma \pm q\sigma_\gamma\right).$$

The validity of the intervals follows immediately from the definition of q. When the covariance matrix of the estimates is not known exactly but can be estimated, one can similarly construct intervals by computing the $1-\alpha$ quantile of the maximal t-statistic.

Conditional approaches. Conditional approaches bound the probability of error conditional on selecting a specific target:

$$P\{\theta_\gamma \notin C_\gamma \mid \hat\gamma = \gamma\}.$$

While simultaneous methods ensure validity for arbitrary $\hat\gamma$ chosen from the set Γ, regardless of any further properties of the selection, conditional approaches adapt to the selection at hand. In particular, a crucial step in implementing a conditional correction is tractably characterizing the selection event $\{\hat{\gamma} = \gamma\}$. Prior work has provided such characterizations for a variety of model selection methods, such as the LASSO, forward stepwise, LARS, etc [61, 106, 164]. This adaptivity of conditional approaches often allows them to outperform simultaneous approaches; for example, if a specific target is selected with overwhelming probability, then conditional methods yield confidence intervals that are nearly the same as uncorrected intervals. On the other hand, characterizing the selection event is difficult in general and conditional corrections are available only in certain restrictive problem settings, usually parametric exponential families. Furthermore, since the final guarantees are conditional rather than unconditional, conditional methods can yield large intervals; notably, Goeman and Solari [68] showed that for every conditional method there exists a simultaneous inference method that dominates it in terms of power. Kivaranovic and Leeb [98] showed that conditionally valid intervals, even if tight, have infinite expected length for common selection problems. To fix this issue of enlarged intervals due to conditioning, Andrews et al. [2] introduced a refinement of conditional inference for the problem of inference on the "winner" called the *hybrid* method. The hybrid method begins by constructing simultaneous intervals for all candidate targets of inference. Then, it implements a correction conditional on both the selected target *and* the event that the intervals constructed in the first step cover the target. This strategy offers unconditional guarantees only, but can lead to significant power gains over standard conditional inference.

Another solution to the enlarged intervals due to conditioning which is relevant to this thesis is the idea of randomizing the selection procedure [14, 97, 133, 134, 158, 159, 161]. Notably, the pioneering work in this direction due to Tian and Taylor [159] proves a central limit theorem that asymptotically relates the validity of statistical inferences without selection to their selective counterparts, a result similar in flavor to a result we will present in Section 4.3. However, existing randomization proposals suffer several drawbacks. One is that they give little insight into the tradeoff between confidence interval width and the loss in utility from the additional noise. Another issue is that inference is based on a selective pivot which, unlike in exact conditional approaches, lacks closed-form expressions. As a result, to approximate the pivot, existing work resorts to computationally expensive sampling [159, 161], which is generally infeasible in high dimensions. There are other, computationally-efficient approaches which aim to approximate the pivot [133, 134], although these are only approximate and the general theory applies to restricted classes of selection problems.

4.3 Validity via algorithmic stability

In this section we develop a theoretical framework that delivers provably valid selective inferences by *randomizing* the selection of the target of inference. Specifically, we build on the concept of *algorithmic stability*, in particular its variant with origins in the field of differential privacy [54], to derive selective confidence intervals that are both tractable computationally and powerful statistically.

We provide a valid correction to classical, non-selective confidence intervals simultaneously for all procedures that have the same level of algorithmic stability. Informally, a selection being stable means that it is not too sensitive to the particular realization of the data, and the more stable the selection is, the smaller the resulting intervals are. In particular, if the selection is "perfectly stable" in the sense that the inferential target is fixed up front and does not depend on the data at hand, the confidence intervals resulting from our approach smoothly recover classical confidence intervals.

Before diving into formal details, we sketch our main result. For simplicity, suppose there is a single inferential target of interest $\hat{\gamma}$, selected in a data-driven way. Imagine that there is an oracle that guesses $\hat{\gamma}$, only knowing the method used to arrive at the selection together with the distribution of the data, but not its realization. Denote by $\hat{\gamma}_0$ the oracle's guess. We say that a selection procedure is η-stable for some $\eta > 0$ if there exists an oracle such that, with high probability over the distribution of the data, the likelihood of any selection under $\hat{\gamma}$ and the likelihood of the same selection under $\hat{\gamma}_0$ can differ by at most a multiplicative factor of e^{η}. Intuitively, η quantifies how much the selection can vary across different realizations of the data; $\eta = 0$ essentially means that the selection cannot depend on the data and hence $\hat{\gamma}$ is fixed, while as η grows the selection is allowed to be increasingly data-adaptive. Note that the magnitude of stability depends not only on the selection method, but also on the distribution of the data.

Our main result provides a post-selection-valid correction to classical, non-selective confidence intervals for stable selection procedures. In short, it says that it suffices to perform standard, uncorrected inference at error level $\alpha e^{-\eta}$ if the goal is to have the final error be at most α, as long as the selection of $\hat{\gamma}$ is η-stable. Therefore, the correction is very simple: it merely says that one should discount the target error level as a function of the selection's stability. In the rest of the section we formalize this statement and explain how stability can be achieved.

4.3.1 Definition of algorithmic stability

The formal theory of algorithmic stability characterizes how the output of an algorithm changes when the input is perturbed. Randomized algorithms have as output a *random variable*; therefore, to study the stability of a randomized algorithm, an appropriate notion of closeness of two random variables is required. The particular notion of closeness considered in differential privacy and related work is known as *indistinguishability*, or *max-divergence.*

Definition 4.3.1 (Indistinguishability). *We say that a random variable Q is (η, τ)-indistinguishable from W, denoted $Q \approx_{\eta,\tau} W$, if for all measurable sets $\mathcal{O}$,*

$$P\{Q \in \mathcal{O}\} \leqslant e^{\eta} P\{W \in \mathcal{O}\} + \tau.$$

Note that indistinguishability is essentially a property of two distributions; for this reason, we will sometimes say that a distribution P_Q is (η, τ)-indistinguishable from a distribution P_W, meaning that $Q \approx_{\eta,\tau} W$ holds for any $Q \sim P_Q$ and $W \sim P_W$.

Roughly speaking, τ bounds the probability of the event where Q and W are "very different." For fixed $\tau \in [0, 1]$, the parameter η is meant to capture how similar the distributions of Q and W are—the larger η is the larger the divergence between Q and W can be. One should think of τ as being at most a small factor proportional to the miscoverage level α.

We now formally introduce the main notion of algorithmic stability considered in this section. The algorithm whose stability we analyze will usually be a selection algorithm. Intuitively, a randomized algorithm $\mathcal{A}$ is stable if there exists an "oracle" random variable A_0 such that, for all "typical" inputs ω, $\mathcal{A}(\omega)$ is distributionally indistinguishable from A_0. In other words, as long as the input is typical, we can approximate the distribution of the randomized algorithm's output with a *fixed* law, without having to see the input in the first place.

Definition 4.3.2 (Stability). *Let $\mathcal{A} : \mathbb{R}^n \to \mathcal{S}$ be a randomized algorithm. We say that $\mathcal{A}$ is (η, τ, ν)-stable with respect to a distribution P supported on $\mathbb{R}^n$ if there exists a random variable A_0, possibly dependent on P, such that*

$$P\left\{\omega \in \mathbb{R}^n : \mathcal{A}(\omega) \approx_{\eta,\tau} A_0\right\} \geqslant 1 - \nu.$$

This notion is a special case of *typical stability* introduced by Bassily and Freund [7]. It is closely related to the notions of perfect generalization [42] and max-information [53]. Unless stated otherwise, whenever we use the term stability we will assume stability in the sense of Definition 4.3.2. The parameter ν can in principle take on any value in $[0, 1]$ but in practice we will set it to be proportional to α.

We will only invoke stability with respect to the data distribution, which we will denote by P_y. Thus, for simplicity, when we say that $\mathcal{A}$ is (η, τ, ν)-stable we are implicitly assuming that it is stable with respect to P_y.

Definition 4.3.2 requires that, as the input data ω varies, the distribution of $\mathcal{A}(\omega)$ remains indistinguishable from a *fixed* distribution that does not depend on ω, namely the distribution of A_0. The parameter ν allows the laws of $\mathcal{A}(\omega)$ and A_0 to deviate for a small set of atypical data vectors ω. The parameters η and τ bound the maximum deviation of $\mathcal{A}(\omega)$ from A_0 over the typical set of vectors ω.

Given a stable algorithm, we will refer to A_0 (which must exist by definition) as its corresponding oracle. The term "oracle" is motivated by the fact that A_0 will typically depend on P_y, which is unknown. To build further intuition, suppose that we observe data $y \sim P_y$ and let $\mu = \mathbb{E}\, y$. Most of our stability constructions will rely on arguing that Definition 4.3.2 holds if we take $A_0 = \mathcal{A}(\mu)$; the reader should think of this as the most prototypical oracle construction. In other words, $\mathcal{A}(y)$ *conditional* on y is indistinguishable from $A(\mu)$ in the sense of Definition 4.3.1 (as long as y is not an atypical data set). At a high level, this happens because y concentrates around μ; we work out a concrete example building on this idea below.

Example 4.3.1. *To provide intuition for Definition 4.3.2, we present one simple mechanism for achieving stability. Although basic, this mechanism will be a fundamental building block* *in our stability proofs. Suppose that we wish to compute $w^\top y$, for some fixed vector w, and suppose that we take P_y to be $\mathcal{N}(\mu, \sigma^2 I)$ with known $\sigma > 0$. Let $\mathcal{A}(y) = w^\top y + \xi$, where $\xi \sim Lap\left(\frac{z_{1-\nu/2}\sigma\|w\|_2}{\eta}\right)$, for user-specified parameters $\eta > 0, \nu \in (0,1)$. Here, $Lap(b)$ denotes a draw from the zero-mean Laplace distribution with parameter b, independent of y. We argue that this mechanism is $(\eta, 0, \nu)$-stable. First, we know*

$$P\{|w^\top y - w^\top \mu| \geqslant z_{1-\nu/2}\sigma\|w\|_2\} = P\{|\mathcal{N}(0, \sigma^2\|w\|_2^2)| \geqslant z_{1-\nu/2}\sigma\|w\|_2\} = \nu.$$

Denote $E = \{\omega \in \mathbb{R}^n : |w^\top \omega - w^\top \mu| \leqslant z_{1-\nu/2}\sigma\|w\|_2\}$, and notice that we have shown that $P\{y \in E\} = 1 - \nu$.

Now let $A_0 = \mathcal{A}(\mu)$. Since the ratio of densities of $\xi \sim Lap(b)$ and its shifted counterpart $x + \xi$ is upper bounded by $e^{|x|/b}$, we can conclude that for all $\omega \in E$ and measurable sets $\mathcal{O}$,

$$\frac{P\{\mathcal{A}(\omega) \in \mathcal{O}\}}{P\{\mathcal{A}(\mu) \in \mathcal{O}\}} \leqslant e^{\eta};$$

that is, we have $\mathcal{A}(\omega) \approx_{\eta,0} A_0$ for all $\omega \in E$. Putting everything together, we see that $\mathcal{A}(\cdot)$ is $(\eta, 0, \nu)$-stable with respect to P_y.

4.3.2 Confidence intervals after stable selection

Given the assumption of (η, τ, ν)-stability, we now show how a simple modification to classical confidence intervals suffices to correct for selective inferences. This correction is valid *regardless* of any additional property of the selection criterion.

The main intuition behind this assertion is the following. If the selection algorithm is stable, then by Definition 4.3.2 one can construct an oracle selection $\hat{\Gamma}_0$ *without looking at* y, such that the actual selected targets $\hat{\Gamma}(y)$ and $\hat{\Gamma}_0$ are distributionally indistinguishable. Since $\hat{\Gamma}(y)$ is indistinguishable from $\hat{\Gamma}_0$, we can pretend that $\hat{\Gamma}_0$ is the selection of interest. Furthermore, since $\hat{\Gamma}_0$ was constructed independently of y, we are *free to use y for inference.* Stability ensures that, despite data reuse, inference behaves almost like with data splitting, in which we perform selection on one batch of data and then use independent data for constructing intervals.

We state a technical lemma, similar to Lemma 3.3 by Bassily and Freund [7], that we use to prove our main theorem.

Lemma 4.3.1. *Let $\hat{\Gamma} : \mathbb{R}^n \to \mathcal{S}$ be an (η, τ, ν)-stable selection algorithm and let $\hat{\Gamma}_0$ be the corresponding oracle selection. Then, it holds that*

$$(y, \hat{\Gamma}(y)) \approx_{\eta,\tau+\nu} (y, \hat{\Gamma}_0). \tag{4.1}$$

Equipped with Lemma 4.3.1, we can now describe how to construct post-selection-valid confidence intervals after stable selection.

Suppose that, under selection Γ', our target of inference is $\{\theta_\gamma\}_{\gamma\in\Gamma'}$. Moreover, suppose that $C^{\alpha}_{\gamma\cdot\Gamma'}$ are valid confidence intervals at level $1-\alpha$ for any *fixed* Γ', meaning that

$$P\{\exists\gamma \in \Gamma' : \theta_\gamma \notin C^{\alpha}_{\gamma\cdot\Gamma'}\} \leqslant \alpha.$$

Such intervals are provided by classical theory.

Theorem 4.3.1 formally states how to construct confidence intervals for an *adaptive* target $\widehat{\Gamma}$, when $\hat{\Gamma}$ is selected in a stable way. This is the key result of this section.

Theorem 4.3.1. *Fix $\delta \in (0,1)$, and let $\hat{\Gamma}$ be an (η,τ,ν)-stable selection algorithm. Then,*

$$P\{\exists\gamma \in \widehat{\Gamma} : \theta_\gamma \notin C^{\delta e^{-\eta}}_{\gamma\cdot\hat{\Gamma}}\} \leqslant \delta+\tau+\nu.$$

In words, if $\hat{\Gamma}$ is (η,τ,ν)-stable, we can pretend that there is no selection bias and simply construct classical intervals, albeit at a more conservative level, to achieve validity. If we set the target error level to be $\delta e^{-\eta}$, then the realized error level will be at most $\delta+\tau+\nu$. For example, if we let $\tau=\nu=\alpha/3$, then to get coverage at level $1-\alpha$ we can set the target coverage level to be $\alpha/3 \cdot e^{-\eta}$.

Comparison with data splitting

In many scenarios it is possible to split the data into two independent chunks, one to be used for selection and the other to be reserved for inference. Classical inferences are then valid because the inferential target is determined before seeing any of the data used in the inference step. This simple baseline for valid inference after selection is called *data splitting.* We illuminate the relationship between our approach via stability and data splitting.

First we want to emphasize that the stability principle is applicable even with dependent samples: Theorem 4.3.1 can be applied even when it is not clear how to create two independent subsets of the data. Moreover, in some selection problems data splitting makes little conceptual sense, such as in our first motivating vignette about inference on the winning effect.

The appeal of data splitting lies in its broad applicability. As long as the data can be split into two independent components, the criteria for choosing the inferential target can be arbitrary. Therefore, data splitting provides a *selection-agnostic* correction, universally valid across all possible selection strategies.

Conceptually, stability lies somewhere between data splitting and conditional post-selection inference. It computes a correction level as a function of how adaptive the selection is to the data, thereby adapting to some properties of the selection rule like conditional inference methods. However, at the same time it provides a correction that is universally valid across all possible selection strategies *with the same level of stability*, which can be seen as a refinement of the principle of data splitting.

To illustrate the conceptual difference between the stability principle and the data splitting principle, suppose that in the latter case we allocate f-fraction of the data to selection, and $(1-f)$-fraction to inference. Then, the resulting intervals will roughly look like classical intervals augmented by a factor of $\sqrt{\frac{1}{1-f}}$ *regardless* of how the selection is performed.

In contrast, the stability approach augments classical intervals as a function of the adaptivity of the selection algorithm. Suppose for concreteness that $y \sim \mathcal{N}(\mu, I)$ and we are considering doing inference on one of two targets, $v_0^\top \mu$ or $v_1^\top \mu$, where the selection $\hat{\gamma} \in \{0,1\}$ depends on the data y. Consider three different selection methods:

- $\hat{\gamma} = 1$ no matter what the data vector is.
- $\hat{\gamma} = 1$ if $\bar{y} := \frac{1}{n}\sum_{i=1}^n y_i \geqslant 0$, and $\hat{\gamma} = 0$ otherwise.
- $\hat{\gamma} = 1$ if $X_1^\top y \geqslant 0$ for some unit vector X_1, and $\hat{\gamma} = 0$ otherwise.

We can write all three procedures as $\hat{\gamma} = \mathbf{1}\{w^\top y \geqslant 0\}$; in the first case $w = 0$, in the second case $w = \frac{1}{n}\mathbf{1}$, and in the third case $w = X_1$.

Let us fix the noise level $\gamma > 0$ and select $\hat{\gamma} = \mathbf{1}\{w^\top y + \xi \geqslant 0\}$, where $\xi \sim \text{Lap}(\gamma)$. The first method is trivially $(0,0,0)$-stable for any level γ, hence we can simply use y for inference without any correction. Based on the same analysis as in Example 4.3.1, the second selection method is $(\sqrt{2\log(2/\nu)}/(\gamma\sqrt{n}), 0, \nu)$-stable for all $\nu > 0$; i.e., it is $(\sqrt{2\log(4/\alpha)}/(\gamma\sqrt{n}), 0, \alpha/2)$-stable. Similarly, the third selection method is $(\sqrt{2\log(4/\alpha)}/\gamma, 0, \alpha/2)$-stable.

We can thus observe that, even though in all three examples we perturb the selection by the same constant level of noise, the stability approach exploits the fact that some selection criteria are more stable than others and this is reflected in the resulting stability parameter. By Theorem 4.3.1, this stability parameter, in turn, directly determines the correction factor, i.e., how conservative we need to make classical inferences for them to be valid post selection.

While data splitting and stability come with conceptual differences, they also have technical similarities. In particular, each one has a leading parameter—$f \in (0,1)$ in the case of data splitting and $\eta > 0$ in the case of stability—and this parameter interpolates between two extremes. One extreme is when all information is reserved for inference (attained when $f = 0$ and $\eta = 0$ respectively) and the other is when all information is used for selection (attained when $f = 1$ and $\eta \to \infty$ respectively). Therefore, it might make sense to ask how the two interpolations relate.

For every η, there is an $f(\eta)$ such that, if we used $f(\eta)$-fraction of the data for selection and $1 - f(\eta)$ for inference, we would approximately get the same interval correction. We sketch the derivation of $f(\eta)$ in the case of normal intervals for simplicity, however this calculation can be generalized to other distributions. We will assume that $\nu + \tau \leqslant \delta\alpha$ for some $\delta \in (0,1)$; then, the intervals resulting from (η, τ, ν)-stability are of width proportional to $z_{1-(1-\delta)\frac{\alpha}{2}e^{-\eta}}$. The intervals resulting from data splitting are of width proportional to $z_{1-\frac{\alpha}{2}}(1 - f(\eta))^{-1/2}$. By equating the two expressions to achieve the same width and simplifying, we obtain

$$f(\eta) = 1 - \left(\frac{z_{1-\frac{\alpha}{2}}}{z_{1-(1-\delta)\frac{\alpha}{2}e^{-\eta}}}\right)^2 \approx \frac{\log\frac{1}{1-\delta} + \eta}{\log\frac{2}{(1-\delta)\alpha} + \eta}, \tag{4.2}$$

where the approximation on the right-hand side follows by a subgaussian approximation.

Of course, this sketch only gives intuition for when data splitting and stability imply equally powerful inference; it does not say anything about which selection is more accurate—one where we select on $f(\eta)$-fraction of the data, or one where we select on the whole data set in an η-stable way. We will tackle this question empirically, as the notion of "more accurate" varies greatly depending on the context.

Finally, we mention another proposal that is conceptually closely related to data splitting, namely the (U, V) decomposition of Rasines and Young [142]. Like stability, the (U, V) decomposition allows the statistician to see all data points—more precisely, noisy versions thereof—both in the selection step and in the inference step. This is an important advantage over data splitting when there are only a few samples that capture information about certain directions. In contrast with stability, performing the (U, V) decomposition does not rely on any properties of the selection method. However, finite-sample guarantees of this approach crucially rely on the data being Gaussian with known covariance, while the stability principle is applicable beyond Gaussianity and is robust to only having an estimate of the covariance.

4.3.3 Model selection in linear regression

In this section, we discuss an application of our stability tools to the problem of model selection in linear regression. We focus on the framework presented in the seminal work of Berk et al. [12], which we reviewed in Section 4.1.1.

The confidence intervals resulting from our approach take the usual form,

$$C_{j\cdot\hat{M}}(K) := \left(\hat{\theta}_{j\cdot\hat{M}} \pm K\hat{\sigma}_{j\cdot\hat{M}}\right),$$

where $\hat{\sigma}^2_{j\cdot\hat{M}}$ is an estimator of variance for the OLS estimate $\hat{\theta}_{j\cdot\hat{M}}$; e.g., the "sandwich" variance estimator [22]. Our goal is to find a suitable value of K such that $C_{j\cdot\hat{M}}(K)$ are valid $(1-\alpha)$-confidence intervals:

$$P\{\theta_{j\cdot\hat{M}} \in C_{j\cdot\hat{M}}(K),\ \forall j \in \hat{M}\} \geqslant 1-\alpha.$$

By analogy with Berk et al. [12], we refer to the minimal such valid K as the *PoSI constant*. It is important to remember that, unlike in Berk et al., our PoSI constant depends on the selection procedure, rather than a family of all possible models.

The PoSI constant is well characterized when the model is fixed rather than determined in a data-driven fashion. For a fixed model M and given $\alpha \in (0,1)$, we define $K_{M,\alpha}$ to be the minimum value of K such that

$$P\left\{\max_{j\in M}\left|\frac{\hat{\theta}_{j\cdot M}-\theta_{j\cdot M}}{\hat{\sigma}_{j\cdot M}}\right| \geqslant K\right\} \leqslant \alpha.$$

In other words, $K_{M,\alpha}$ defines the PoSI constant when the model M is specified up front and does not depend on the data; in this case, $C_{j\cdot M}(K_{M,\alpha})$ are valid simultaneous intervals at level $1-\alpha$. For example, when $y \sim \mathcal{N}(\mu, \sigma^2 I)$, one simple way of providing a valid upper bound on $K_{M,\alpha}$ is via standard z-scores or t-scores, after doing a Bonferroni correction over $j \in M$. Sharper estimates of $K_{M,\alpha}$ can be obtained by exploiting the correlations between the regression coefficients to estimate the maximum z-score or t-score; see Section 4.2.

We are now ready to state a corollary of Theorem 4.3.1 that focuses on the problem of model selection in linear regression.

Corollary 4.3.1. *Fix $\delta \in (0,1)$. Let $\hat{M}$ be an (η,τ,ν)-stable model selection algorithm. For all $j \in \hat{M}$, let:*

$$C_{j\cdot\hat{M}}(K_{\hat{M},\delta e^{-\eta}}) = \left(\hat{\theta}_{j\cdot\hat{M}} \pm K_{\hat{M},\delta e^{-\eta}}\hat{\sigma}_{j\cdot\hat{M}}\right).$$

Then,

$$P\left\{\exists j \in \hat{M} : \theta_{j\cdot\hat{M}} \notin C_{j\cdot\hat{M}}\left(K_{\hat{M},\delta e^{-\eta}}\right)\right\} \leqslant \delta + \tau + \nu.$$

To provide further intuition, we instantiate Corollary 4.3.1 in the canonical setting of Gaussian observations. Let $y \sim \mathcal{N}(\mu, \sigma^2 I)$. If $\sigma > 0$ is known, we let $\hat{\sigma}_{j\cdot M} = \sigma\sqrt{((X_M^\top X_M)^{-1})_{jj}}$; otherwise, we assume we have access to an estimate of σ, denoted $\hat{\sigma}$, and let $\hat{\sigma}_{j\cdot M} = \hat{\sigma}\sqrt{((X_M^\top X_M)^{-1})_{jj}}$. Following the treatment of Berk et al. [12], we assume that $\hat{\sigma}^2 \sim \sigma^2\frac{\chi_r^2}{r}$ for r degrees of freedom and assume that $\hat{\sigma}^2 \perp \hat{\theta}_{j\cdot M}$ for all possible OLS estimates $\hat{\theta}_{j\cdot M}$. If the full model is assumed to be correct, that is $y \sim \mathcal{N}(X\theta, \sigma^2 I)$, and $n > d$, then this assumption is satisfied for $r = n-d$ by setting $\hat{\sigma}^2 = \|y - X\hat{\theta}\|_2^2/(n-d)$, where $\hat{\theta}$ is the OLS estimate in the full model. Even if the full model is not correct, there exist other ways of producing such a valid estimate of σ; we refer the reader to Berk et al. [12] for further discussion.

We denote by $z_{1-\alpha}$ the $1-\alpha$ quantile of the standard normal distribution, and by $t_{r,1-\alpha}$ the $1-\alpha$ quantile of the t-distribution with r degrees of freedom.

Corollary 4.3.2. *Fix $\delta \in (0,1)$, and suppose $y \sim \mathcal{N}(\mu, \sigma^2 I)$. Further, let $\hat{M}$ be an (η,τ,ν)-stable model selection algorithm. If σ is known, let:*

$$C_{j\cdot\hat{M}} = \left(\hat{\theta}_{j\cdot\hat{M}} \pm z_{1-\delta/(2|\hat{M}|e^{\eta})}\sigma\sqrt{((X_{\hat{M}}^\top X_{\hat{M}})^{-1})_{jj}}\right).$$

If, on the other hand, σ is not known but there exists an estimate, $\hat{\sigma}^2 \sim \sigma^2\frac{\chi_r^2}{r}$, independent of the OLS estimates, let:

$$C_{j\cdot\hat{M}} = \left(\hat{\theta}_{j\cdot\hat{M}} \pm t_{r,1-\delta/(2|\hat{M}|e^{\eta})}\hat{\sigma}\sqrt{((X_{\hat{M}}^\top X_{\hat{M}})^{-1})_{jj}}\right).$$

In either case, we have

$$P\{\exists j \in \hat{M} : \theta_{j\cdot\hat{M}} \notin C_{j\cdot\hat{M}}\} \leqslant \delta + \tau + \nu.$$

The proof follows by a direct application of Corollary 4.3.1, together with a Bonferroni correction over $j \in \hat{M}$ when computing $K_{\hat{M},\delta e^{-\eta}}$. Approximating Gaussian quantiles by subgaussian concentration, we observe that the PoSI constant in Corollary 4.3.2 scales roughly as $\sqrt{2\left(\log(2|\hat{M}|/\delta) + \eta\right)}$ (when σ is known, or as $r \to \infty$ when σ is estimated from data).

Recovering the Scheffé rate

Our main technical step in deriving selective confidence intervals is Lemma 4.3.1, which argues that the joint distribution of $(y, \hat{\Gamma})$ cannot be too different from the joint distribution of $(y, \hat{\Gamma}_0)$, where $\hat{\Gamma}_0$ is the oracle from the definition of stability, in the indistinguishability metric. In the context of model selection in linear regression, we verify that the confidence intervals resulting from this approach are not vacuously wide in the two most extreme settings: the first, in which the model selection is independent of the data, and the second, in which the model selection is arbitrarily complex and dependent on the data.

Suppose that $\hat{M}$ is independent of y. Then, the distribution of $\hat{M}(y)$, conditional on y, is equal to the distribution of $\hat{M}(\omega)$ for any point ω, hence $\hat{M}(\omega)$ is an oracle which trivially implies $(0,0,0)$-stability. In this case, the intervals in Corollary 4.3.1 reduce to $C_{j\cdot\hat{M}}(K_{\hat{M},\delta})$ and are valid at level $1-\delta$, as expected.

Now suppose that $\hat{M}$ is allowed to have arbitrary dependence on y; in particular, it can attain the "significant triviality bound" of Berk et al. [12]. While arguing stability in the sense of Definition 4.3.2 would require additional assumptions, the only property of stability used to prove Theorem 4.3.1—the indistinguishability bound in Eq. (4.1)—can be obtained. This allows for the proof of Theorem 4.3.1 to go through, thus recovering the tight rate of existing analyses.

Proposition 4.3.1. *Let $\hat{M}$ be an arbitrary, possibly randomized model selection procedure, such that $|\hat{M}| \leqslant s$ almost surely. Then, for any P_y, there exists an oracle selection $\hat{M}_0$ such that for any $\tau \in (0,1)$,*

$$(y, \hat{M}(y)) \approx_{\eta,\tau} (y, \hat{M}_0), \text{ for some } \eta = O(s\log(d/s)) + \log(1/\tau).$$

Consequently, there exists a value $\eta = O(s\log(d/s)) + \log(1/\tau)$ such that the confidence intervals $C_{j\cdot\hat{M}}(K_{\hat{M},\delta e^{-\eta}}) = \left(\hat{\theta}_{j\cdot\hat{M}} \pm K_{\hat{M},\delta e^{-\eta}}\hat{\sigma}_{j\cdot\hat{M}}\right)$ satisfy

$$P\left\{\exists j \in \hat{M} : \theta_{j\cdot\hat{M}} \notin C_{j\cdot\hat{M}}\left(K_{\hat{M},\delta e^{-\eta}}\right)\right\} \leqslant \delta + \tau.$$

By approximating Gaussian quantiles via subgaussian concentration, we obtain confidence intervals which are universally valid for *all* s-sparse selections under Gaussian outcomes and scale as $O(\sqrt{\eta}) = O(\sqrt{s\log(d/s)})$. This rate is in general tight [102], and as s approaches d, it matches the rate given by the Scheffé protection [12, 149].

4.3.4 The design of stable selection algorithms

We discuss general tools for designing stable selection methods and present an application of these tools to variable selection in linear regression. We begin with an overview of the basic properties of stability, which are key to efficient design of stable selections.

Properties of stability

Stability satisfies two key algorithmic properties: *closure under post-processing* and *composition.* We provide precise definitions of the two shortly. The reason why these properties enable efficient stability designs is that many selection rules can be written as post-processing and composition of simple computations, such as linear functions of the data or finding maxima of a sequence. As long as we know how to stabilize the necessary simple computations, closure under post-processing and composition provide rules for computing the overall stability parameter of the whole algorithm efficiently.

Post-processing. First, stability is *closed under post-processing*: if $\mathcal{A} : \mathbb{R}^n \to \mathcal{S}$ is (η, τ, ν)-stable, then for any (possibly randomized) map $\mathcal{B} : \mathcal{S} \to \mathcal{G}$, the composition $\mathcal{B} \circ \mathcal{A}$ is also (η, τ, ν)-stable. While the proof of this fact is a straightforward consequence of the definition of stability, the implications are significant. Suppose for the moment that the statistician is given a stable version of the LASSO algorithm, and denote its solution by $\hat{\beta}_{\text{LASSO}}$. Since $\hat{\beta}_{\text{LASSO}}$ is stable, then so is

$$\hat{M} = \{j \in [d] \ : \ \hat{\beta}_{\text{LASSO},j} \neq 0\}.$$

In fact, the statistician need not necessarily choose the model corresponding *exactly* to the support of $\hat{\beta}_{\text{LASSO}}$; for example, they could choose $\hat{M} = \{j \in [d] : |\hat{\beta}_{\text{LASSO},j}| \geqslant \varepsilon\}$, for some constant threshold ε, or they could pick $d_{\text{sel}} \leqslant d$ entries with the maximum absolute value. More generally, any model chosen solely as a function of $\hat{\beta}_{\text{LASSO}}$ inherits the same stability parameters as $\hat{\beta}_{\text{LASSO}}$. And, according to Corollary 4.3.1, the same PoSI constant suffices to correct the confidence intervals resulting from any such model.

Composition. The second important property is *composition.* In Algorithm 4, we define adaptive composition, after which we discuss simpler, non-adaptive composition.

Algorithm 4 Adaptive composition

```
input: data $y \in \mathbb{R}^n$, sequence of algorithms $\mathcal{A}_t : \mathcal{S}_1 \times \cdots \times \mathcal{S}_{t-1} \times \mathbb{R}^n \to \mathcal{S}_t,\ t \in [k]$
output: $(a_1, \dots, a_k) \in \mathcal{S}_1 \times \cdots \times \mathcal{S}_k$
  for $t = 1, 2, \dots, k$ do
    Compute $a_t = \mathcal{A}_t(a_1, \dots, a_{t-1}, y) \in \mathcal{S}_t$
  end for
  Return $(a_1, \dots, a_k)$
```

Adaptive composition consists of k sequential rounds in which the analyst observes the outcomes of all previous computations and selects the next computation *adaptively*—as a function of the previous evaluations. The adaptive composition property bounds the stability parameters of Algorithm 4 in terms of the stability parameters of $\mathcal{A}_t$. In its simplest form, it says that Algorithm 4 is $(k\eta, 0, 0)$-stable if for all $t \in [k]$, $\mathcal{A}_t(a_1, \dots, a_{t-1}, \cdot)$ is $(\eta, 0, 0)$-stable for all fixed $a_1, \dots, a_{t-1}$. For example, for some selection algorithms such as forward stepwise, it is clear to see how they can be represented using adaptive composition. In forward stepwise, $\mathcal{A}_t$ outputs an index $i_t \in [d]$, which corresponds to the variable i that minimizes the squared error resulting from adding i to the current pool of selected features; $i_t = \mathcal{A}_t(i_1, \dots, i_{t-1}, y)$. It suffices to prove that any given step of forward stepwise selection is stable, in order to infer that the overall algorithm is stable as well. More generally, greedy algorithms can naturally be represented using adaptive composition (see [69] for an application in the context of greedy causal discovery algorithms).

Our proofs will only require adaptive composition for algorithms with $\nu = 0$; such results follow from classical theory on differential privacy. More advanced (and naturally more conservative) adaptive composition theorems which allow $\nu > 0$ can be found in the context of typical stability [7].

A simpler kind of composition is non-adaptive composition. Here, the algorithms $\mathcal{A}_t$ have no dependence on the past computations. Non-adaptive composition can capture a protocol that involves running multiple selection methods and choosing a final selection target as an arbitrary function of all the outputs. The resulting stability parameters simply add up. This is a rather appealing property of stability, as it suggests that the statistician only needs to keep track of the stability parameters of each selection algorithm they run, in order to derive valid selective confidence intervals. An analogous combination of the results of different selection methods was considered by Markovic and Taylor [118]; their approach, however, relies on a sophisticated Monte Carlo sampling scheme.

Model selection algorithms: examples

We now consider several algorithms for variable selection in linear regression through the lens of stability. While many of the principles presented in this section can be adapted to different distributional assumptions, for the sake of clarity and interpretability we assume that $y \sim \mathcal{N}(\mu, \sigma^2 I)$, where σ^2 is *unknown* but we have access to an estimate $\hat{\sigma}^2 \sim \sigma^2 \frac{\chi_r^2}{r}$, independent of y. This is the setup studied by Berk et al. [12]. More generally, we only need to know the decay of the tail of the distribution of y in order to enforce stability. For example, we can handle outcome vectors with a known bound on their Orlicz norm, for any Orlicz function. This includes general subgaussian and subexponential outcome vectors.

Model selection via the LASSO. We begin by considering the canonical example of the LASSO estimator [162]. The LASSO estimate is the solution to the usual least-squares problem with an additional ℓ_1-constraint on the regression coefficients:

$$\hat{\beta}_{\text{LASSO}} \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2}\|y - X\theta\|_2^2 \text{ s.t. } \|\theta\|_1 \leqslant C_1, \tag{4.3}$$

where $C_1 > 0$ is a tuning parameter. This problem is sometimes referred to as the LASSO in constrained/bound form, to contrast it with the LASSO in penalized form:

$$\hat{\beta}^{\lambda}_{\text{LASSO}} \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2}\|y - X\theta\|_2^2 + \lambda\|\theta\|_1, \tag{4.4}$$

where $\lambda > 0$ is now the tuning parameter. These two problems are equivalent in a sense: given X and y, for any $C_1 > 0$ there exists a corresponding $\lambda > 0$ such that $\hat{\beta}_{\text{LASSO}}$ is an optimal solution for the problem in Eq. (4.4), and vice versa. In our analysis we focus on the formulation (4.3). It is worth pointing out that our selective inference tools do not directly extend to penalized LASSO since, for a fixed penalty λ, the corresponding constraint C_1 depends on the data, which is random. Extending our approach to handle inference after solving the penalized problem is an important direction for future work.

The LASSO objective induces sparse solutions, and a common way of declaring that a feature is relevant is to check for a corresponding non-zero entry in the LASSO solution vector. That is, the model "selected" by the LASSO is:

$$\hat{M} = \{j \in [d] \ : \ \hat{\beta}_{\text{LASSO},j} \neq 0\}.$$

Model selection via the LASSO has been of great interest in prior work on selective inference, starting with Lee et al. [106]. While this work provides exact confidence intervals, it has been observed that these intervals (which do not make use of randomization) have infinite expected length [98]. Subsequent work has improved upon these often large confidence intervals by applying randomization [97, 133, 134, 158, 159, 161].

We now formulate a stable version of the LASSO algorithm. It is inspired by the differentially private LASSO algorithm of Talwar et al. [157], although the noise variables are calibrated somewhat differently due to different modeling assumptions.

We use e_i to denote the i-th standard basis vector in $\mathbb{R}^d$, and $\{\pm e_i\}_{i=1}^d$ to denote the set of $2d$ standard basis vectors, multiplied by 1 and -1. We also let $\|X\|_{2,\infty}$ denote the $L_{2,\infty}$ norm of X, $\|X\|_{2,\infty} := \max_{i\in[d]} \|X_i\|_2$.

Algorithm 5 Stable LASSO algorithm

```
input: design matrix $X \in \mathbb{R}^{n\times d}$, outcome vector $y \in \mathbb{R}^n$, variance estimate $\hat{\sigma}^2 \sim \sigma^2\frac{\chi_r^2}{r}$, $\ell_1$-constraint $C_1$, number of steps $k$, parameters $\delta \in (0,1), \eta > 0$
output: LASSO solution $\hat{\beta}_{\text{LASSO}} \in \mathbb{R}^d$
  Initialize $\beta_1 = 0$
  for $t = 1, 2, \dots, k$ do
    $\forall\phi \in C_1 \cdot \{\pm e_i\}_{i=1}^d$, sample $\xi_{t,\phi} \overset{\text{i.i.d.}}{\sim} \text{Lap}\left(\frac{4t_{r,1-\delta/(2d)}C_1\|X\|_{2,\infty}}{\eta n}\right)$
    $\forall\phi \in C_1 \cdot \{\pm e_i\}_{i=1}^d$, let $\alpha_\phi = -\frac{2}{n\hat{\sigma}}\phi^\top X^\top(y - X\beta_t) + \xi_{t,\phi}$
    Set $\phi_t = \arg\min_{\phi\in C_1\cdot\{\pm e_i\}_{i=1}^d} \alpha_\phi$
    Set $\beta_{t+1} = (1-\Delta_t)\beta_t + \Delta_t\phi_t$, where $\Delta_t = \frac{2}{t+1}$
  end for
  Return $\hat{\beta}_{\text{LASSO}} = \beta_{k+1}$
```

In essence, Algorithm 5 is a randomized version of the classical Frank-Wolfe algorithm from constrained optimization [63].

We now argue that $\hat{\beta}_{\text{LASSO}}$ is stable. The proof is based on a composition argument: namely, we can view $\hat{\beta}_{\text{LASSO}}$ as the result of a composition of k subroutines, each given by one optimization step which produces β_t. The stability of each subroutine is proved by extending an argument related to the "report noisy max" mechanism from differential privacy [55].

Proposition 4.3.2 (LASSO stability)**.** *Algorithm 5 is both*

(a) $\left(\frac{1}{2}k\eta^2 + \sqrt{2k\log(1/\delta)}\eta, \delta, \delta\right)$*-stable, and*

(b) $(k\eta, 0, \delta)$*-stable.*

We state two rates because there exist parameter regimes where either rate leads to tighter confidence intervals than the other (the first rate being tighter when η is small).

By the post-processing property, Proposition 4.3.2 implies stability of any model $\hat{M}$ obtained as a function of $\hat{\beta}_{\text{LASSO}}$, such as the model corresponding to its non-zero entries.

Notice that the noise level in Algorithm 5 is an explicit function of η. This allows the statistician to understand the loss in utility—that is, how much worse $\hat{\beta}_{\text{LASSO}}$ is relative to an exact LASSO solution—due to randomization. In fact, building on work by Jaggi [89] and Talwar et al. [157], we can upper bound the excess risk resulting from randomization.

Proposition 4.3.3 (LASSO utility)**.** *Suppose we run Algorithm 5 for* $k = \left\lceil \frac{n\|X\|_\infty^2 C_1\eta}{\hat{\sigma}\|X\|_{2,\infty}} \right\rceil$ *steps. Then,*

$$\frac{1}{n}\mathbb{E}[\|y - X\hat{\beta}_{\text{LASSO}}\|_2^2 \mid y] - \min_{\beta:\|\beta\|_1 \leqslant C_1} \frac{1}{n}\|y - X\beta\|_2^2 = \tilde{O}\left(\frac{C_1\|X\|_{2,\infty}\log(d)t_{r,1-\delta/(2d)}\sigma}{n\eta}\right).$$

Model selection via marginal screening. One of the most commonly used model selection methods involves simply picking a constant number of the features with the largest absolute inner product with the outcome y [60, 73]. That is, one selects features i corresponding to the top k values of $|X_i^\top y|$, for a pre-specified parameter k. This strategy is known as *marginal screening*, and it was first analyzed in the context of selective inference by Lee and Taylor [107].

In Algorithm 6, we state a stable version of marginal screening. Notice that the randomization scheme is similar to that of the stable LASSO method. As before, we let $\|X\|_{2,\infty}$ denote the $L_{2,\infty}$ norm of X.

Algorithm 6 Stable marginal screening algorithm

```
input: design matrix $X \in \mathbb{R}^{n\times d}$, outcome vector $y \in \mathbb{R}^n$, variance estimate $\hat\sigma^2 \sim \sigma^2 \frac{\chi_r^2}{r}$, model size $k$, parameters $\delta \in (0,1), \eta > 0$
output: $\hat M = \{i_1, \ldots, i_k\}$
  Compute $(c_1, \ldots, c_d) = \frac{1}{n\hat\sigma} X^\top y \in \mathbb{R}^d$
  $\text{res}_1 = [d]$
  for $t = 1, 2, \ldots, k$ do
    $\forall i \in \text{res}_i$, sample $\xi_{t,i} \overset{\text{i.i.d.}}{\sim} \text{Lap}\left(\frac{2t_{r,1-\delta/(2d)}\|X\|_{2,\infty}}{n\eta}\right)$
    $i_t = \arg\max_{i\in\text{res}_t} |c_i + \xi_{t,i}|$
    $\text{res}_{t+1} = \text{res}_t \setminus i_t$
  end for
  Return $\hat M = \{i_1, \ldots, i_k\}$
```

The high-level idea behind the proof of stability of Algorithm 6 is similar to that of Algorithm 5.

Proposition 4.3.4 (Marginal screening stability)**.** *Algorithm 6 is both*

(a) $\left(\frac{1}{2}k\eta^2 + \sqrt{2k\log(1/\delta)}\eta, \delta, \delta\right)$*-stable, and*

(b) $(k\eta, 0, \delta)$*-stable.*

As for the LASSO, we aim to quantify the loss in utility due to randomization. Given that the goal of marginal screening is to detect the largest k values $|c_i| = |X_i^\top y|$, a reasonable notion of utility loss is the difference between the values c_i corresponding to the variables in $\hat M$, and the actual largest values of c_i.

Proposition 4.3.5 (Marginal screening utility)**.** *Let m_i denote the index of the i-th largest value c_j in absolute value, so that $(|c_{m_1}|, \ldots, |c_{m_d}|)$ is the decreasing order statistic of $\{|c_i|\}_{i=1}^d$. Then, for any $\delta' \in (0,1)$, Algorithm 6 satisfies:*

$$P\left\{\max_{j\in[k]} |c_{m_j}| - |c_{i_j}| \leqslant \frac{4t_{r,1-\delta/(2d)}\log(dk/\delta')\|X\|_{2,\infty}}{n\eta}\Big| y\right\} \geqslant 1 - \delta'.$$

4.3.5 Experimental results

In this section, we evaluate our selective intervals for the LASSO and marginal screening and compare our solution with data splitting.

For a fixed sample size n we vary the number of features d. We consider two different data-generating processes for the design matrix: one in which the rows of X are drawn independently from an equicorrelated multivariate Gaussian distribution with pairwise correlation $\rho = 0.5$, and the second one in which all entries of X are drawn as independent

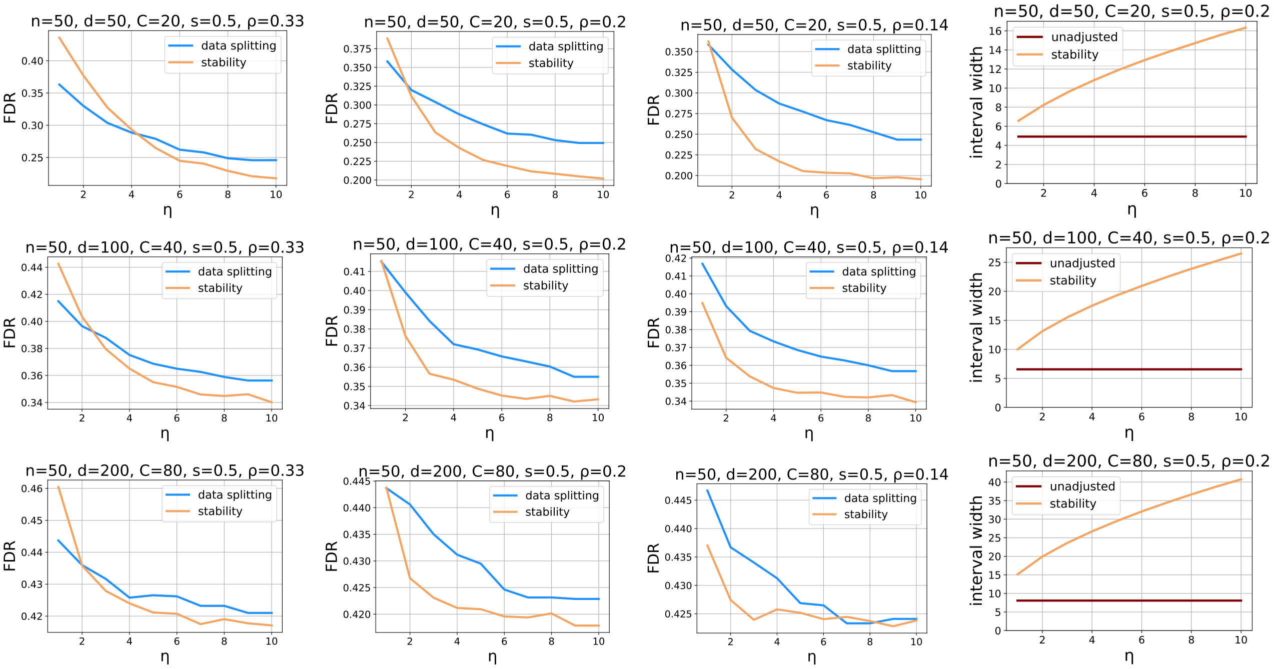


Figure 4.1: Comparison of FDR after stable LASSO and LASSO with data splitting, with varying dimension and signal strength, in the Gaussian design case. In addition, we plot the average interval width (at $\rho = 0.2$ only, however the width varies minimally with ρ) and the average unadjusted width.

Bernoulli random variables with parameter 0.1. In the former case, X is normalized to have columns of unit norm. The outcome is generated as $y = X\beta + \varepsilon$, where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1), i \in [n]$,

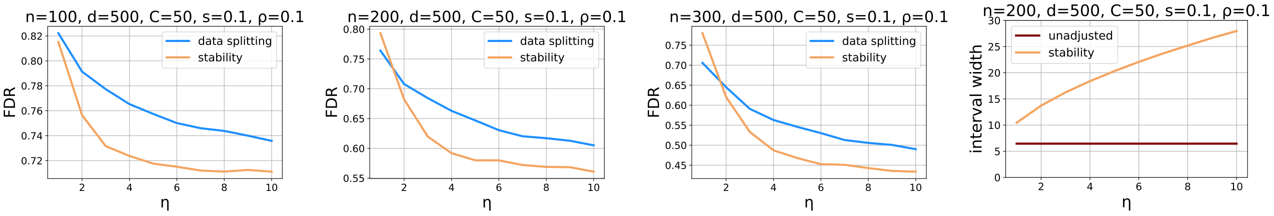


Figure 4.2: Comparison of FDR after stable LASSO and LASSO with data splitting, with varying sample size, in the Gaussian design case. In addition, we plot the average interval width at $n = 200$ and the average unadjusted width.

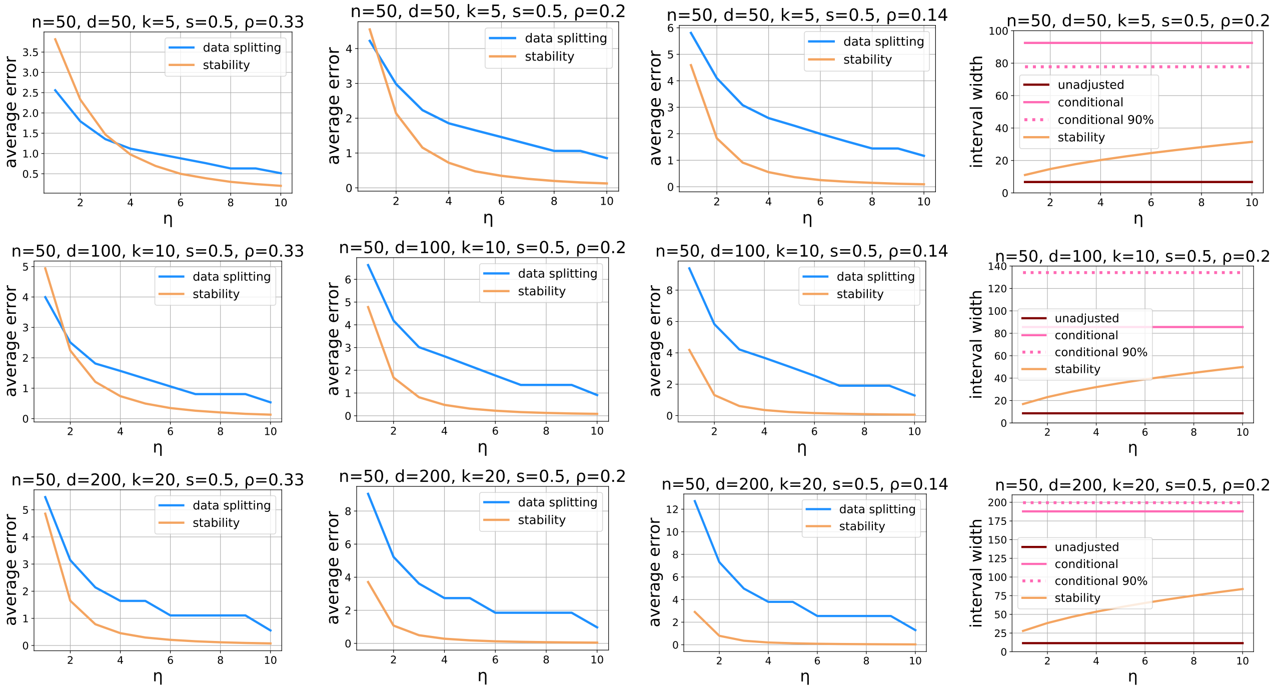


Figure 4.3: Comparison of average error after stable marginal screening and marginal screening with data splitting, with varying dimension and signal strength, in the Gaussian design case. In addition, we plot the average interval width (at $\rho = 0.2$ only, however the width varies minimally with ρ), together with the average unadjusted width and the width obtained via the conditional correction of Lee and Taylor [107]. We also plot the 90% quantile of the conditional width because it varies greatly across realizations.

and the entries of β are sampled according to

$$\beta_i = \begin{cases} \mathrm{Exp}(\rho), \; i \in \{1, \dots, sd\}, \\ 0, \; i \in \{sd+1, \dots, d\}, \end{cases}$$

for a signal parameter $\rho > 0$ and a sparsity parameter $s \in (0, 1)$, which we vary. We fix the target miscoverage level to be $\alpha = 0.1$. In all experiments we vary $\eta \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. For the comparison with data splitting, we use the splitting fraction derived in Section 4.3.2.

Gaussian design

We first state the results for the Gaussian design case.

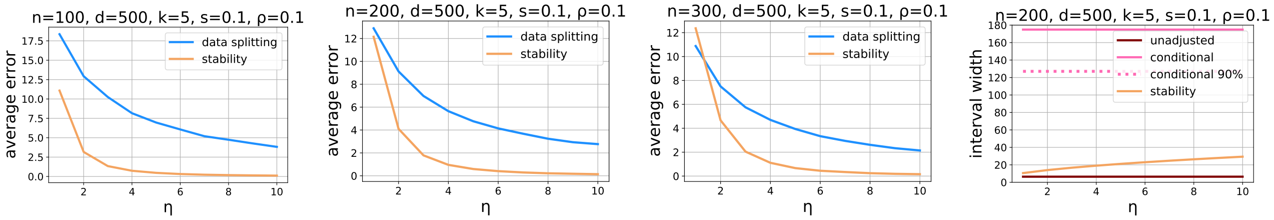


Figure 4.4: Comparison of average error after stable marginal screening and marginal screening with data splitting, with varying sample size, in the Gaussian design case. In addition, we plot the average interval width at $n = 200$, together with the average unadjusted width and the width implied by the conditional approach of Lee and Taylor [107].

LASSO. In Figure 4.1 we compare the false discovery rate (FDR) of the stable LASSO algorithm and the LASSO algorithm with data splitting. In all plots $n = 50$ is fixed and we vary $d \in \{50, 100, 200\}$. As we increase d, we also increase the size of the constraint set $C_1 \in \{20, 40, 80\}$ to allow more selections. We consider signal levels $\rho \in \{0.33, 0.2, 0.14\}$, which corresponds to an expected value of the non-null β_i lying in $\{3, 5, 7\}$, and we fix $s = 0.5$.

We observe that stability generally outperforms data splitting as η grows, equivalently when the splitting fraction $f(\eta)$ grows, as well as when the signal strength grows. In Figure 4.1 we additionally plot the average width of stable intervals against the average width of naive, unadjusted intervals. Note that the intervals obtained via data splitting have essentially the same width (and are hence not plotted), based on how $f(\eta)$ is chosen. We only plot interval width for $\rho = 0.2$ since the width varies minimally for different values of ρ.

In Figure 4.2 we compare the stable LASSO algorithm and the LASSO with data splitting in a sparse high-dimensional setting with $d = 500, s = 0.1$, and we vary the sample size $n \in \{100, 200, 300\}$. We fix $\rho = 0.1$. We observe that stability consistently outperforms data splitting for large enough η and this gap grows with n. In addition, we plot the average interval width implied by stability against the average unadjusted interval width at $n = 200$ (again we do not plot the interval width given by data splitting for the same reason as in Figure 4.1).

Marginal screening. In Figure 4.3 we compare the average error of stable marginal screening and marginal screening with data splitting. Since marginal screening explicitly aims to maximize the values $|X_i^\top y|$ for selected variables X_i, we quantify the error as $\frac{1}{k}\sum_{t=1}^{k}(|X_{i_t^*}^\top y| - |X_{i_t}^\top y|)$, where i_t is the estimated index of the t-th largest absolute inner product (based on a subsample in the case of data splitting, or based on a randomized sample in the case of stability), and i_t^* is the true index of the t-th largest absolute inner product in the data set. We vary the parameters as in the LASSO comparison in Figure 4.1, only instead of varying C_1 we vary $k \in \{5, 10, 20\}$. We also plot the average interval width

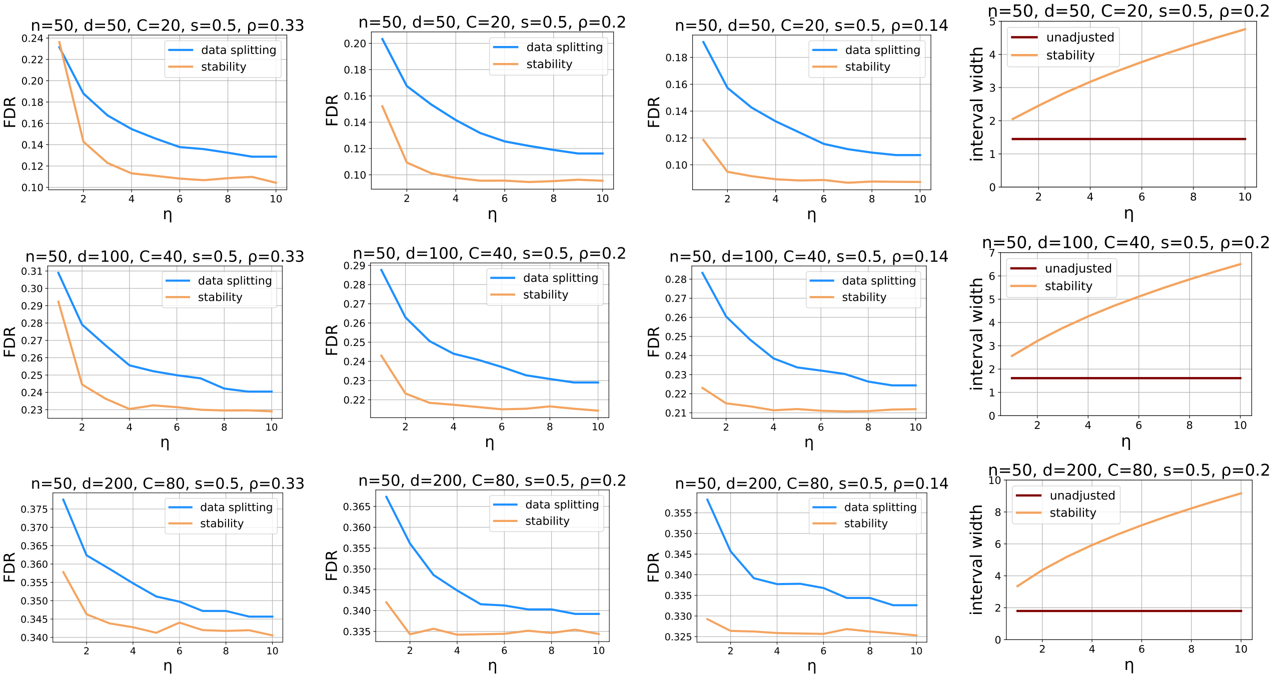


Figure 4.5: Comparison of FDR after stable LASSO and LASSO with data splitting, with varying dimension and signal strength, in the Bernoulli design case. In addition, we plot the average interval width (at $\rho = 0.2$ only, however the width varies minimally with ρ) and the average unadjusted width.

with stability, together with the unadjusted interval width and the average width obtained via the conditional method of Lee and Taylor [107] with no randomization. For the conditional method, since the intervals are sometimes orders of magnitude larger than the average width, we also plot the 90% quantile of interval width. We see that stability typically out-

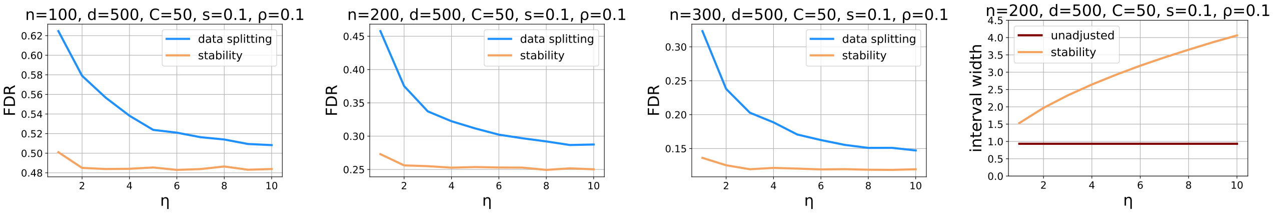


Figure 4.6: Comparison of FDR after stable LASSO and LASSO with data splitting, with varying sample size, in the Bernoulli design case. In addition, we plot the average interval width at $n = 200$ and the average unadjusted width.

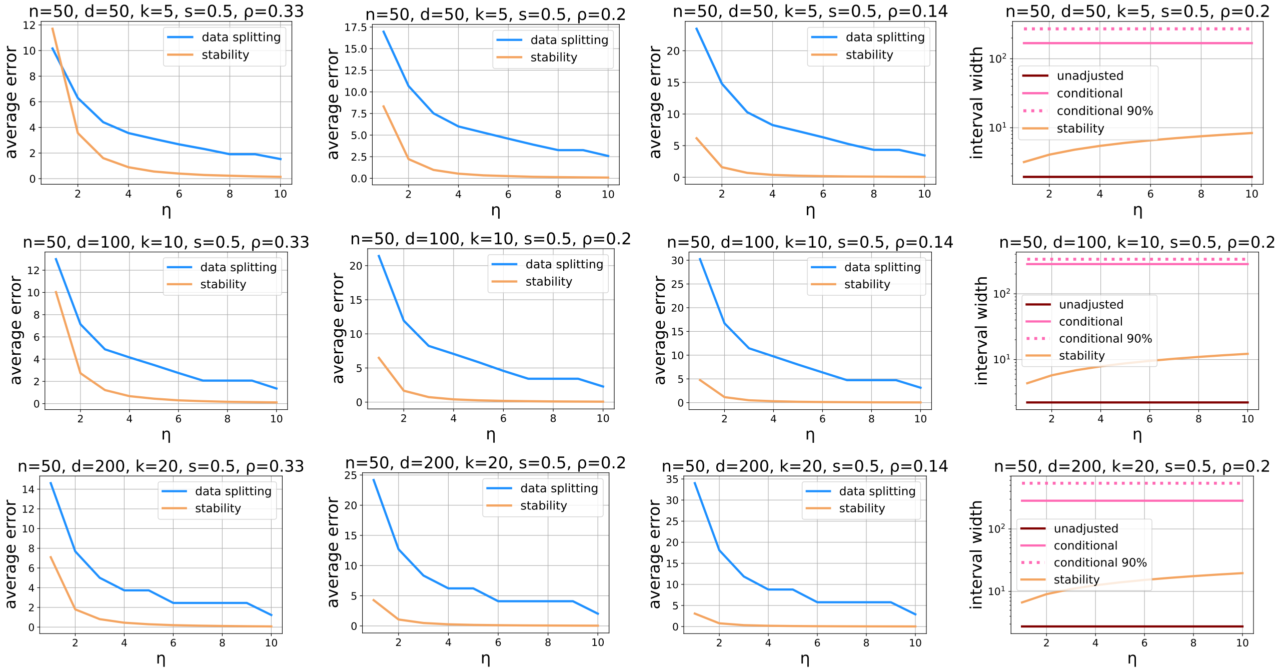


Figure 4.7: Comparison of average error after stable marginal screening and marginal screening with data splitting, with varying dimension and signal strength, in the Bernoulli design case. In addition, we plot the average interval width (at $\rho = 0.2$ only, however the width varies minimally with ρ), together with the average unadjusted width and the width obtained via the conditional correction of Lee and Taylor [107]. We also plot the 90% quantile of the conditional width because it varies greatly across realizations. Since the conditional widths are of a higher order of magnitude, the scale on the y-axis in the widths plots is logarithmic.

performs data splitting in terms of the average error, and this benefit is more pronounced for larger η and signal strength. In terms of interval width, we observe that stability leads to significantly smaller intervals than the conditional approach. We plot interval width when $\rho = 0.2$.

In Figure 4.4 we consider a setting analogous to that of Figure 4.2, and we analogously vary the sample size n. We again see that stability generally dominates data splitting. Moreover, the gap between the intervals obtained via stability and those of Lee and Taylor [107] is even more pronounced than in Figure 4.3.

Bernoulli design

Now we consider the Bernoulli design case. The motivation for considering a sparse Bernoulli design lies in the fact that certain directions in the column space of X are captured by only a few samples, hence missing out on them—as is possible with data splitting—can significantly

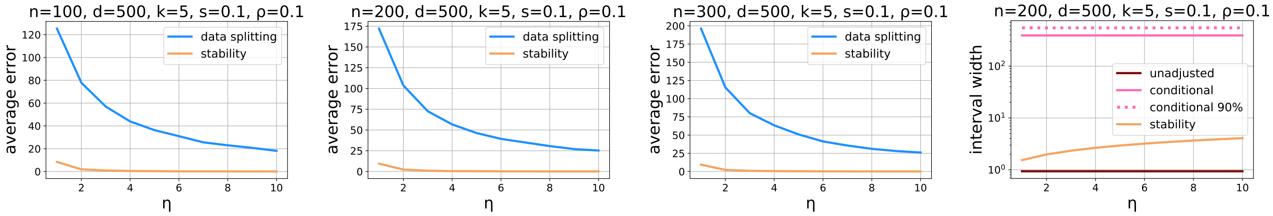


Figure 4.8: Comparison of average error after stable marginal screening and marginal screening with data splitting, with varying sample size, in the Bernoulli design case. In addition, we plot the average interval width at $n = 200$, together with the average unadjusted width and the width implied by the conditional approach of Lee and Taylor [107]. Since the conditional widths are of a higher order of magnitude, the scale on the y-axis in the widths plot is logarithmic.

affect the quality of selection.

LASSO. In Figure 4.5 and Figure 4.6 we provide comparisons analogous to those of Figure 4.1 and Figure 4.2, using the same parameter configurations. We observe a larger gap between data splitting and stability than in the Gaussian design case, and observe the same trends: as η and the signal strength grow, the performance gap increases.

Marginal screening. In Figure 4.7 and Figure 4.8 we provide comparisons analogous to those of Figure 4.3 and Figure 4.4, using the same parameter configurations. We observe a larger gap between data splitting and stability both than in the Gaussian design case, as well as in the LASSO experiments using the Bernoulli design. In addition, we observe an even more pronounced gap between stable confidence interval widths and widths of intervals obtained via a conditional correction [107]. For this reason, the y-axis in the widths plots is logarithmic.

4.4 Validity via locally simultaneous inference

Simultaneous inference is still one of the most common strategies for ensuring valid selective inference. It is broadly applicable, robust to parametric assumptions, and often amenable to efficient implementation. However, simultaneous inference can be unnecessarily conservative when many questions, although possible, are unlikely to be of interest in the first place. For example, suppose that a clinical trial estimates the effectiveness of multiple treatments and, after observing the data, it is clear that there are many ineffective treatments and only a handful of effective ones. Even if we are only interested in constructing a confidence interval for the effectiveness of the best-performing treatment, simultaneous inference would still widen the intervals enough to cover all possible treatments, including the clearly ineffective ones that never stood a chance of being selected. Similarly, if we use a method like the LASSO to select a sparse subset of variables for a linear model, it may be clear in hindsight that most variables had no chance of beind selected. In both of these cases, simultaneous inference can be very conservative.

In this section we introduce *locally simultaneous inference*, an approach that ensures valid selective inference while only answering those questions that could *plausibly* have been selected in light of the observed data. Locally simultaneous inference comes with rigorous type I error guarantees like simultaneous inference but is less conservative; in particular, it reduces to simultaneous inference in an extreme case. In the clinical trial example, locally simultaneous inference would require taking a correction only over reasonably effective treatments, while testing the ineffective ones comes virtually for free.

To sketch our main idea, suppose that we have a family of estimands $\{\theta_\gamma : \gamma \in \Gamma\}$, where Γ indexes all admissible targets of inference. In the running example, Γ would be the set $\{1, \dots, m\}$, where m is the total number of treatments in the clinical trial, and θ_γ is the mean effect of the indexed treatment. Given data y, we are interested in doing inference on $\theta_{\hat{\gamma}}$, where $\hat{\gamma}$ is a data-dependent target chosen from Γ; in the running example, $\hat{\gamma}$ indexes the treatment that seems most effective according to y. It would be invalid to reuse the same data y to perform an uncorrected inference on $\theta_{\hat{\gamma}}$, since the winning treatment is likely to have been overestimated by the trial data. However, if it is clear in hindsight that only a small number $k \ll m$ of the treatments were even in the running to win, it would be wasteful to make the full multiplicity correction for all m treatments.

The main idea behind our framework is to find a *data-dependent* set of targets $\widehat{\Gamma}^+$, which is nested between the selected target and all possible targets, $\hat{\gamma} \in \widehat{\Gamma}^+ \subseteq \Gamma$, such that taking a standard simultaneous correction over $\widehat{\Gamma}^+$ ensures valid selective inferences. Perhaps surprisingly, this strategy is valid despite the dependence between $\widehat{\Gamma}^+$ and the data. Moreover, if the selection $\hat{\gamma}$ is "obvious enough" in hindsight, $\widehat{\Gamma}^+$ only contains $\hat{\gamma}$ and our approach nearly reduces to classical, uncorrected inference.

Unlike simultaneous inference, our approach adapts to the specifics of the selection criterion; in this sense, locally simultaneous inference resembles *conditional* selective inference, which delivers valid inference after conditioning on the event that a specific target was selected. However, since our approach builds on the robust and broadly applicable principle of simultaneous inference, it comes with several advantages over conditional inference, including numerical stability and robustness to parametric assumptions.

To give a glimpse of the comparison of locally simultaneous inference to standard simultaneous inference and conditional inference, we consider a simple illustrative example. Suppose that $y_1 \sim \mathcal{N}(\mu_1, 1), y_2 \sim \mathcal{N}(\mu_2, 1)$ are independent and we wish to do inference on the mean of observation $\hat{\gamma} = \arg\max_{\gamma \in \{1,2\}} y_\gamma$. When the gap $\Delta = \mu_2 - \mu_1$ is near zero, the inferential question of interest is most uncertain, while large Δ corresponds to the case where the inferential question of interest is "obvious". In Figure 4.9 we plot the median, together with the 5% and 95% quantile, of the width of selective confidence intervals constructed via locally simultaneous inference, standard simultaneous inference, and conditional inference. We observe that for small Δ conditional inference can lead to large intervals, and as Δ

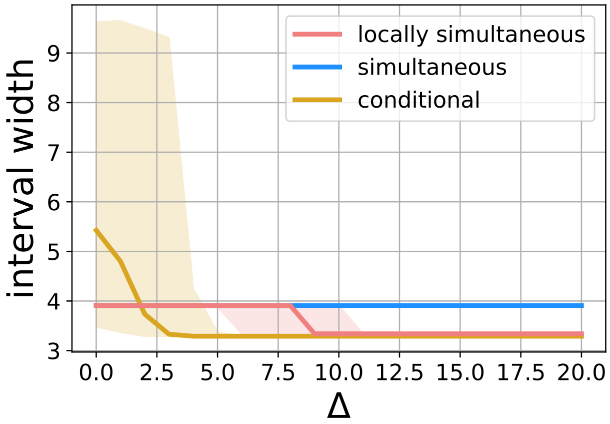


Figure 4.9: Interval width achieved by locally simultaneous inference, fully simultaneous inference, and conditional inference. The data is $(y_1, y_2) \sim \mathcal{N}((\mu_1, \mu_2), I_2)$ and the goal is to do inference on the mean of observation $\hat{\gamma} = \arg\max_{\gamma\in\{1,2\}} y_\gamma$. We vary $\Delta = \mu_2 - \mu_1$.

grows conditional intervals approach nominal, unadjusted intervals. Simultaneous inference is insensitive to the value of Δ and delivers constant-width intervals, which are smaller than conditional intervals for small Δ due to their unconditional nature. Locally simultaneous inference adapts to the certainty of the selection like conditional inference, but is never worse than simultaneous inference. Formally, by relying on our general theory of locally simultaneous inference, we obtain the following approach. Fix $\alpha = 0.1$. Let $q^\delta(k)$ be the $1-\delta$ quantile of $\max_{i\in[k]} |Z_i|, Z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Then, we have

$$P\left\{\mu_{\hat{\gamma}} \in \left(y_{\hat{\gamma}} \pm \min\{q^{0.95\alpha}(|\widehat{\Gamma}^+|), q^\alpha(2)\}\right)\right\} \geqslant 1-\alpha,$$

where $\widehat{\Gamma}^+ = \{1,2\}$ when $|y_2 - y_1| \leqslant 2\sqrt{2}q^{0.05\alpha}(1)$ and $\widehat{\Gamma}^+ = \{\hat{\gamma}\}$ otherwise. Therefore, when $|y_2 - y_1|$ is small, locally simultaneous intervals are equal to simultaneous intervals; when $|y_2 - y_1|$ is large, only $\hat{\gamma}$ is deemed to be a plausible selection in hindsight, making the intervals essentially uncorrected.

4.4.1 General construction

The basic principle of our correction is to find a *data-dependent* set of targets nested between $\widehat{\Gamma}$ and Γ, such that taking a simultaneous correction over the set ensures type I error control. To implement this idea, we assume that we can construct simultaneous confidence regions $C_{\gamma\cdot\Gamma'}$ for any desired subset $\Gamma' \subseteq \Gamma$, at any target error level α. Formally, we have access to a family of confidence regions $\{C_{\gamma\cdot\Gamma'}\}_{\Gamma'\subseteq\Gamma}$ such that

$$P\{\theta_\gamma \in C_{\gamma\cdot\Gamma'},\ \forall\gamma\in\Gamma'\} \geqslant 1-\alpha,$$

for all $\Gamma' \subseteq \Gamma$.

To ensure validity of our construction, we make a mild and natural monotonicity assumption, requiring that the confidence regions can only increase as the set of inferential targets increases.

Assumption 4.4.1. *We say that the confidence regions* $\{C_{\gamma\cdot\Gamma'}\}_{\Gamma'\subseteq\Gamma}$ *are* nested *if for all* $\Gamma_1 \subseteq \Gamma_2 \subseteq \Gamma$ *and* $\gamma \in \Gamma_1$,

$$C_{\gamma\cdot\Gamma_1} \subseteq C_{\gamma\cdot\Gamma_2}.$$

We are now ready to outline the general solution based on locally simultaneous inference.

For every $P \in \mathcal{P}$, suppose that we can construct a set $A_\nu(P)$ that satisfies

$$P\{y \in A_\nu(P)\} \geqslant 1 - \nu,$$

for any pre-specified $\nu \in (0, 1)$. In other words, $A_\nu(P)$ is the acceptance region of a valid test for the null hypothesis $H_P : y \sim P$ at level $1 - \nu$. Intuitively, $A_\nu(P)$ can be thought of as the set of all plausible observations according to distribution P. When $\mathcal{P} = \{P_\mu\}_{\mu\in\mathcal{M}}$ is a parametric family, we will simply write $A_\nu(P_\mu) \equiv A_\nu(\mu)$.

We define the set of *plausible targets* under distribution P to be:

$$\Gamma_\nu(P) := \underset{y'\in A_\nu(P)}{\cup} \widehat{\Gamma}(y').$$

Note that, unlike the realized selection $\widehat{\Gamma}(y)$, $\Gamma_\nu(P)$ is a *fixed* set of targets. Again, when $\mathcal{P} = \{P_\mu\}_{\mu\in\mathcal{M}}$ is a parametric family, we will write $\Gamma_\nu(P_\mu) \equiv \Gamma_\nu(\mu)$.

Finally, we define the inversion of $A_\nu(P)$, which gives a confidence region for the true distribution P:

$$B_\nu(y) = \{P \in \mathcal{P} : y \in A_\nu(P)\}.$$

Before proving our main result, which asserts validity of locally simultaneous inference, we prove a key technical lemma that makes the core of the argument.

Lemma 4.4.1. *Fix* $\alpha \in (0, 1)$ *and* $\nu \in (0, \alpha)$. *Let* $\{\tilde{C}_{\gamma\cdot\Gamma'}\}_{\Gamma'\subseteq\Gamma}$ *be a family of confidence regions such that*

$$P\left\{\theta_\gamma \in \tilde{C}_{\gamma\cdot\Gamma'}, \forall\gamma \in \Gamma', y \in A_\nu(P)\right\} \geqslant 1 - \alpha, \tag{4.5}$$

for all $\Gamma' \subseteq \Gamma$. *Moreover, suppose that the regions are nested (Ass. 4.4.1). Consider the set of targets*

$$\widehat{\Gamma}^+_\nu = \underset{P'\in B_\nu(y)}{\cup} \Gamma_\nu(P') = \underset{P'\in B_\nu(y)}{\cup}\ \underset{y'\in A_\nu(P')}{\cup} \widehat{\Gamma}(y').$$

Then, it holds that

$$P\left\{\theta_\gamma \in \tilde{C}_{\gamma\cdot\widehat{\Gamma}^+_\nu}, \forall\gamma \in \widehat{\Gamma}\right\} \geqslant 1 - \alpha.$$

Proof. First, we can write

$$P\left\{\theta_\gamma \in \tilde{C}_{\gamma\cdot\widehat{\Gamma}_\nu^+}, \forall\gamma \in \widehat{\Gamma}\right\} \geqslant P\left\{\theta_\gamma \in \tilde{C}_{\gamma\cdot\widehat{\Gamma}_\nu^+}, \forall\gamma \in \widehat{\Gamma}, y \in A_\nu\right\}.$$

Now notice that, on the event $\{y \in A_\nu(P)\} = \{P \in B_\nu(y)\}$, it almost surely holds that $\widehat{\Gamma} \subseteq \Gamma_\nu(P) \subseteq \widehat{\Gamma}_\nu^+$. Using this fact, we have

$$\begin{aligned} P\left\{\theta_\gamma \in \tilde{C}_{\gamma\cdot\widehat{\Gamma}_\nu^+}, \forall\gamma \in \widehat{\Gamma}, y \in A_\nu\right\} &\geqslant P\left\{\theta_\gamma \in \tilde{C}_{\gamma\cdot\widehat{\Gamma}_\nu^+}, \forall\gamma \in \Gamma_\nu(P), y \in A_\nu\right\} \\ &\geqslant P\left\{\theta_\gamma \in \tilde{C}_{\gamma\cdot\Gamma_\nu(P)}, \forall\gamma \in \Gamma_\nu(P), y \in A_\nu\right\}, \end{aligned}$$

where the second inequality follows by the nestedness of the confidence regions. Since the right-hand side is at least $1-\alpha$ by the definition of $\tilde{C}_{\gamma\cdot\Gamma_\nu(P)}$, we have shown

$$P\left\{\theta_\gamma \in \tilde{C}_{\gamma\cdot\widehat{\Gamma}_\nu^+}, \forall\gamma \in \widehat{\Gamma}\right\} \geqslant 1-\alpha,$$

as desired. □

Therefore, Lemma 4.4.1 reduces the problem of constructing selective confidence regions to the problem of constructing the regions $\tilde{C}_{\gamma\cdot\Gamma'}$ satisfying Eq. (4.5) for every *fixed* Γ'. The following theorem, providing such regions, states our main result on locally simultaneous inference.

Theorem 4.4.1. *Fix $\alpha \in (0,1)$ and $\nu \in (0,\alpha)$. Suppose that the simultaneous confidence regions $\{C_{\gamma\cdot\Gamma'}\}_{\Gamma'\subseteq\Gamma}$ are nested (Ass. 4.4.1). Consider the set of targets*

$$\widehat{\Gamma}_\nu^+ = \bigcup_{P'\in B_\nu(y)} \Gamma_\nu(P') = \bigcup_{P'\in B_\nu(y)} \bigcup_{y'\in A_\nu(P')} \widehat{\Gamma}(y'). \tag{4.6}$$

Then, it holds that

$$P\left\{\theta_\gamma \in C^{(\alpha-\nu)}_{\gamma\cdot\widehat{\Gamma}_\nu^+}, \forall\gamma \in \widehat{\Gamma}\right\} \geqslant 1-\alpha.$$

Proof. The proof follows by an application of Lemma 4.4.1. In particular, by a union bound it follows that $C^{(\alpha-\nu)}_{\gamma\cdot\Gamma'}$ is a valid choice of $\tilde{C}_{\gamma\cdot\Gamma'}$ in Lemma 4.4.1:

$$\begin{aligned} P\left\{\theta_\gamma \in C^{(\alpha-\nu)}_{\gamma\cdot\Gamma'}, \forall\gamma \in \Gamma', y \in A_\nu(P)\right\} &\geqslant 1 - P\{\exists\gamma \in \Gamma' : \theta_\gamma \notin C^{(\alpha-\nu)}_{\gamma\cdot\Gamma'}\} - P\{y \notin A_\nu(P)\} \\ &\geqslant 1-(\alpha-\nu)-\nu = 1-\alpha. \end{aligned}$$

□

Intuitively, Theorem 4.4.1 justifies the following refinement of simultaneous inference. Given data y, first construct a set of all distributions under which the observed data is plausible. Then, consider all plausible observations under those distributions; this essentially gives a collection of datasets y' in a neighborhood around y. Finally, perform simultaneous inference over all inferential targets $\widehat{\Gamma}(y')$ that could be selected in this neighborhood. Despite the fact that the set of targets is constructed as a function of y, a simultaneous correction over this set nevertheless ensures valid selective inferences for $\widehat{\Gamma}(y)$.

Next, we provide a slightly different correction from that of Theorem 4.4.1 that *strictly dominates* simultaneous inference at error level α, for any choice of ν (note that the correction in Theorem 4.4.1 dominates simultaneous inference at level $\alpha - \nu$). This is achieved by carefully choosing $A_\nu(P)$. The refined correction is often easy to apply, however we find that the strategy from Theorem 4.4.1 is usually more practical as it allows choosing $A_\nu(P)$ freely. For the next result, we assume *centered* confidence intervals.

Assumption 4.4.2. *We say that $\{C^\alpha_{\gamma\cdot\Gamma'}\}_{\gamma'\subseteq\Gamma}$ are* centered *confidence intervals if*

$$C^\alpha_{\gamma\cdot\Gamma'} = \left(\hat{\theta}_\gamma \pm q^\alpha_{\Gamma'} \cdot \hat{\sigma}_\gamma\right),$$

for some estimator $\hat{\theta}_\gamma$ and standard error $\hat{\sigma}_\gamma$, where $q^\alpha_{\Gamma'}$ is chosen such that $P\{\theta_\gamma \in C^\alpha_{\gamma\cdot\Gamma'}, \forall\gamma \in \Gamma'\} \geqslant 1-\alpha$.

Confidence intervals are often centered; for example, this is true of intervals based on the maximal z- or t-statistic, as in Example 4.2.2. We denote $C_\gamma(q) := \left(\hat{\theta}_\gamma \pm q\hat{\sigma}_\gamma\right)$; then, $C_{\gamma\cdot\Gamma'} = C_\gamma(q^\alpha_{\Gamma'})$ are intervals valid simultaneously over Γ' at level $1-\alpha$. Without loss of generality we assume that $q^\alpha_{\Gamma'}$ is nonincreasing in α.

Theorem 4.4.2. *Fix $\alpha \in (0,1)$ and $\nu \in (0,\alpha)$. Suppose that the confidence intervals are nested (Ass. 4.4.1), i.e., $q^\alpha_{\Gamma_1} \leqslant q^\alpha_{\Gamma_2}$ for all $\Gamma_1 \subseteq \Gamma_2$, and centered (Ass. 4.4.2). Let $A_\nu(P) = \{\theta_\gamma \in C_\gamma(q^\nu_\Gamma), \forall\gamma \in \Gamma\}$, and let $\widehat{\Gamma}^+_\nu$ denote the set of targets from Theorem 4.4.1 (Eq. (4.6)). Let*

$$\hat{q} = \min\left\{q^{(\alpha-\nu)}_{\widehat{\Gamma}^+_\nu}, q^\alpha_\Gamma\right\}.$$

Then, it holds that

$$P\left\{\theta_\gamma \in C_\gamma(\hat{q}), \forall\gamma \in \widehat{\Gamma}\right\} \geqslant 1-\alpha.$$

Proof. Analogously to Theorem 4.4.1, we show that $C_\gamma\left(\min\{q^{(\alpha-\nu)}_{\Gamma'}, q^\alpha_\Gamma\}\right)$ is a valid choice of $\tilde{C}_{\gamma\cdot\Gamma'}$ in Lemma 4.4.1. Invoking Lemma 4.4.1 then completes the proof.

We split the analysis into two cases, depending on which term achieves the minimum.

First, suppose that Γ' is such that $q^{(\alpha-\nu)}_{\Gamma'} \leqslant q^\alpha_\Gamma$. Then, by a union bound, we have

$$\begin{aligned} P\{\theta_\gamma \in \tilde{C}_{\gamma\cdot\Gamma'}, \forall\gamma \in \Gamma', y \in A_\nu(P)\} &\geqslant 1 - P\{y \notin A_\nu(P)\} - P\{\exists\gamma \in \Gamma' : \theta_\gamma \notin \tilde{C}_{\gamma\cdot\Gamma'}\} \\ &= 1 - P\{y \notin A_\nu(P)\} - P\{\exists\gamma \in \Gamma' : \theta_\gamma \notin C_\gamma(q^{(\alpha-\nu)}_{\Gamma'})\} \\ &\geqslant 1-\nu-(\alpha-\nu) \\ &= 1-\alpha. \end{aligned}$$

Next, suppose that Γ' is such that $q_\Gamma^\alpha \leqslant q_{\Gamma'}^{(\alpha-\nu)}$. Then,

$$\begin{aligned} P\{\theta_\gamma \in \tilde{C}_{\gamma\cdot\Gamma'}, \forall\gamma\in\Gamma', y\in A_\nu(P)\} &= P\{\theta_\gamma \in C_\gamma(q_\Gamma^\alpha), \forall\gamma\in\Gamma', y\in A_\nu(P)\} \\ &\geqslant P\{\theta_\gamma \in C_\gamma(q_\Gamma^\alpha), \forall\gamma\in\Gamma, y\in A_\nu(P)\} \\ &= P\{\theta_\gamma \in C_\gamma(q_\Gamma^\alpha), \forall\gamma\in\Gamma\} \\ &\geqslant 1-\alpha, \end{aligned}$$

where the third step follows because, by definition, $\{\theta_\gamma \in C_\gamma(q_\Gamma^\alpha), \forall\gamma\in\Gamma\} \Rightarrow \{y \in A_\nu(P)\}$.

Therefore, $\tilde{C}_{\gamma\cdot\Gamma'} = C_\gamma\left(\min\{q_{\Gamma'}^{(\alpha-\nu)}, q_\Gamma^\alpha\}\right)$ is a valid choice of $\tilde{C}_{\gamma\cdot\Gamma'}$ in Lemma 4.4.1, as desired. □

Locally simultaneous inference therefore comes at *no cost* in terms of power: the intervals are at least as tight as fully simultaneous intervals. Moreover, whenever $\widehat{\Gamma}_\nu^+$ is a strict subset of all admissible selections, they will be strictly tighter.

4.4.2 Inference on the "most promising" effects

We first study the problem of constructing confidence intervals for the "most promising" effects. We consider two instantiations of the problem: inference on the winner and the file-drawer problem.

Given data $y = (y_1, \ldots, y_m) \in \mathbb{R}^m$, the problem of inference on the winner asks for a confidence interval for the mean of the largest entry of y. Formally, if we let $\theta_\gamma = \mathbb{E}\, y_\gamma$ for all $\gamma \in [m]$, the goal is to do inference on $\theta_{\hat\gamma}$, where

$$\hat\gamma = \underset{\gamma\in[m]}{\arg\max}\, y_\gamma. \tag{4.7}$$

The file-drawer problem asks for a confidence region that simultaneously covers the means of all observations that exceed a critical threshold T. Formally, the region is required to cover $\{\theta_\gamma : \gamma \in \widehat{\Gamma}\}$, where

$$\widehat{\Gamma} = \{\gamma \in [m] : y_\gamma \geqslant T\}. \tag{4.8}$$

The m coordinates of y can correspond to, for example, the effectiveness of m different treatments, in which case the selection corresponds to focusing on the single seemingly best treatment or multiple treatments that are deemed sufficiently promising. The m outcomes can also correspond to measurements of a time series over m time steps (e.g., blood pressure in a specified interval), in which case selection focuses on the time steps at which the series achieves extreme values. Finally, y can capture an estimate of the effectiveness of a treatment on m different subgroups (e.g., m age groups); the selection would then ask for the effectiveness within the single subgroup or several subgroups for which the treatment seems most promising.

We consider a parametric version and a nonparametric version of the two problems. Importantly, conditional selective inference is not directly applicable in the latter setting.

Parametric case

We begin with the case where $\mathcal{P}$ is a parametric family; in particular, we take $\mathcal{P} = \{P_\mu\}_\mu$ to be a location family with location parameter $\mu \in \mathbb{R}^m$. In other words, $y = (y_1, \ldots, y_m) \sim P_\mu$ can be written as $y = \mu + Z$, where $Z = (Z_1, \ldots, Z_m) \sim P_0$. For simplicity of exposition we assume that the errors Z_i have the same marginal symmetric zero-mean distribution $P_0^{(1)}$ (e.g., $P_0^{(1)} = \mathcal{N}(0, \sigma^2)$), however generalizing beyond this setting is straightforward. We do not assume that the errors Z_i are necessarily independent, i.e. that P_0 is a product distribution.

For an index set $\mathcal{I} \subseteq [m]$, we define

$$q^\alpha(\mathcal{I}) = \inf\left\{q : P_0\left\{\max_{i\in\mathcal{I}} |Z_i| \leqslant q\right\} \geqslant 1-\alpha\right\}.$$

In words, $q^\alpha(\mathcal{I})$ is the $1-\alpha$ quantile of the maximum absolute error over indices in $\mathcal{I}$. This would be the usual interval half-width if a simultaneous correction is required over the observations in $\mathcal{I}$. This value can be loosely upper bounded by taking a Bonferroni correction over $\mathcal{I}$. We note that exact knowledge of P_0 is not necessary; being able to compute an upper bound on $q^\alpha(\mathcal{I})$ suffices.

Note that $\theta_i = \mathbb{E}\, y_i \equiv \mu_i$ in this setting; that is, the possible estimands θ_i are coordinates of the location parameter. Therefore, for $\hat{\gamma}$ as in Eq. (4.7), we want to construct a confidence interval for $\mu_{\hat{\gamma}}$; for $\widehat{\Gamma}$ as in Eq. (4.8), we want to construct a confidence region for $\{\mu_\gamma : \gamma \in \widehat{\Gamma}\}$.

We now apply our general result about locally simultaneous inference.

Theorem 4.4.3. *Fix $\alpha \in (0,1)$ and $\nu \in (0, \alpha)$.*

- *For the problem of inference on the winner (Eq. (4.7)), let the set of plausible indices be*
$$\widehat{\Gamma}_\nu^+ = \{\gamma \in [m] : y_\gamma \geqslant y_{\hat{\gamma}} - 4q^\nu([m])\}.$$
Then,
$$P_\mu\left\{\mu_{\hat{\gamma}} \in \left(y_{\hat{\gamma}} \pm \min\left\{q^{(\alpha-\nu)}(\widehat{\Gamma}_\nu^+), q^\alpha([m])\right\}\right)\right\} \geqslant 1-\alpha.$$

- *For the file-drawer problem (Eq. (4.8)), let the set of plausible indices be*
$$\widehat{\Gamma}_\nu^+ = \{\gamma \in [m] : y_\gamma \geqslant T - 2q^\nu([m])\}.$$
Then,
$$P_\mu\left\{\mu_\gamma \in \left(y_\gamma \pm \min\left\{q^{(\alpha-\nu)}(\widehat{\Gamma}_\nu^+), q^\alpha([m])\right\}\right),\ \forall \gamma \in \widehat{\Gamma}\right\} \geqslant 1-\alpha.$$

Theorem 4.4.3 formalizes the intuition that one should only have to add the "nearly selected" observations to the simultaneous correction, if the goal is to construct a valid confidence region around the selected ones. When there are many observations that are far from promising, then $\widehat{\Gamma}_\nu^+$ can be much smaller than $[m]$.

We note the file-drawer problem asks for a confidence region around *all* parameters that exceed the selection threshold; in contrast, the conditional approach of Lee et al. [106] provides inference for one real-valued parameter at a time. It is unclear how to generalize it to the problem of inference on multiple parameters without resorting to a trivial solution such as a Bonferroni correction over all estimands, which ignores the dependencies between the different estimation problems. The same observation applies to the hybrid method of Andrews et al. [2]. In contrast, since locally simultaneous inference builds on standard simultaneous inference, it is able to adapt to the dependencies at hand.

Nonparametric case

We show that essentially the same reasoning as in the parametric case applies to nonparametric settings.

For each of the m candidates, we assume that we have n i.i.d. observations that are bounded in $[0,1]$. More formally, we observe n i.i.d. samples $y^{(1)},\dots,y^{(n)}$ drawn from a distribution P with $\mathrm{supp}(P)\subseteq[0,1]^m$. As before, we denote the m-dimensional vector of means by $\theta=\mathbb{E}\,y^{(1)}$.

In the problem of inference on the winner, we would like to do inference on $\theta_{\hat\gamma}$, where

$$\hat\gamma=\arg\max_{\gamma\in[m]} y_\gamma := \arg\max_{\gamma\in[m]} \frac{1}{n}\sum_{j=1}^n y_\gamma^{(j)}. \tag{4.9}$$

In the file-drawer problem, we would like to do inference on $\{\theta_\gamma:\gamma\in\widehat\Gamma\}$, where

$$\widehat\Gamma=\{\gamma\in[m]:y_\gamma\geqslant T\}:=\left\{\gamma\in[m]:\frac{1}{n}\sum_{j=1}^n y_\gamma^{(j)}\geqslant T\right\}. \tag{4.10}$$

Let w_n^α be any valid bound on the deviation of the empirical average of n i.i.d. random variables $X_1,\dots,X_n\in[0,1]$ from their mean. Formally, w_n^α satisfies

$$P\left\{\mathbb{E}\,X_1\in\left(\frac{1}{n}\sum_{i=1}^n X_i\pm w_n^\alpha\right)\right\}\geqslant 1-\alpha.$$

For example, a standard choice of w_n^α is obtained from Hoeffding's inequality:

$$w_n^\alpha=\sqrt{\frac{\log(2/\alpha)}{2n}}.$$

Tighter choices of w_n^α are generally possible, e.g. by applying Bentkus' [11], Bernstein's [13], or Bennett's inequality [10]. Furthermore, for every $\gamma\in[m]$, we let C_γ^α be a confidence region for θ_γ valid at level $1-\alpha$. In our nonparametric experiments, we will take C_γ^α to be the betting-based confidence intervals due to Waudby-Smith and Ramdas [177].

Similarly as in the parametric setting, we ensure valid selective inference by only requiring simultaneous control—here achieved by taking a Bonferroni correction—over the selected and nearly selected observations.

Theorem 4.4.4. *Fix $\alpha \in (0,1)$ and $\nu \in (0,\alpha)$. Assume that $C_\gamma^{\alpha_1} \supseteq C_\gamma^{\alpha_2}$ for all $\alpha_1, \alpha_2 \in (0,1)$ such that $\alpha_1 \leqslant \alpha_2$.*

- *For the problem of inference on the winner (Eq. (4.9)), let the set of plausible indices be*
$$\widehat{\Gamma}_\nu^+ = \left\{\gamma \in [m] : y_\gamma \geqslant y_{\hat{\gamma}} - 4w_n^{\nu/m}\right\}.$$
Then,
$$P\left\{\theta_{\hat{\gamma}} \in C_{\hat{\gamma}}^{(\alpha-\nu)/|\widehat{\Gamma}_\nu^+|}\right\} \geqslant 1-\alpha.$$

- *For the file-drawer problem (Eq. (4.10)), let the set of plausible indices be*
$$\widehat{\Gamma}_\nu^+ = \left\{\gamma \in [m] : y_\gamma \geqslant T - 2w_n^{\nu/m}\right\}.$$
Then,
$$P\left\{\theta_\gamma \in C_\gamma^{(\alpha-\nu)/|\widehat{\Gamma}_\nu^+|},\ \forall \gamma \in \widehat{\Gamma}\right\} \geqslant 1-\alpha.$$

4.4.3 Inference after model selection

We next consider the problem of inference after data-driven model selection. We first state a general implication of locally simultaneous inference in this context and then specialize this result to selection via the LASSO.

Suppose that we have a fixed design matrix $X \in \mathbb{R}^{n\times d}$ and a corresponding vector of outcomes $y \in \mathbb{R}^n$, where $y \sim P_\mu$. We assume P_μ is a location family, that is, $y \sim P_\mu \Leftrightarrow y \overset{d}{=} \mu + Z$, where $Z \sim P_0$ has mean zero.

We want to select a *model* $\hat{M} \equiv \hat{M}(y)$ corresponding to a subset of the d features, and then regress the outcome onto the selected features. We define $\theta_M, \hat{\theta}_M$, etc as Section 4.3.

The set of possible estimands in this context is all possible values of $\theta_{j\cdot M}$. The natural index set Γ for these estimands is given by all feature–model pairs: $\Gamma = \{(j,M) : j \in M, M \in \mathcal{M}\}$, where $\mathcal{M}$ corresponds to all admissible feature selections, which is often $2^{[d]}$. The selected targets are the regression coefficients in the selected model, i.e. $\widehat{\Gamma} = \{(j,\hat{M}) : j \in \hat{M}\}$.

To apply a locally simultaneous correction, we need to compute the augmented set of targets $\widehat{\Gamma}_\nu^+$. The key step in doing so is to find *all plausible models*, which we will denote by $\widehat{\mathcal{M}}_\nu^+$; after we have $\widehat{\mathcal{M}}_\nu^+$, we apply a simultaneous correction in the vein of Berk et al. [12], called the *PoSI* correction, over $\widehat{\mathcal{M}}_\nu^+$. Intuitively, $\widehat{\mathcal{M}}_\nu^+$ is the set of all models that could be selected on outcome vectors similar to y. Again, we note that this set is data-dependent; Berk et al., on the other hand, consider a deterministic set of possible models.

We will focus on methods that use $X^\top y$ as a sufficient statistic, which includes most common selection methods such as the LASSO, forward stepwise, etc. For such methods, a natural choice for the plausible set $A_\nu(\mu)$ is outcome vectors for which $X^\top y \approx X^\top \mu$. We formalize this in Corollary 4.4.1 below.

To state the result, for a set of contrasts $\mathcal{V}$, we define

$$q^{\alpha}(\mathcal{V}) = \inf \left\{ q : P_0 \left\{ \sup_{v \in \mathcal{V}} |v^\top Z| \leqslant q \right\} \geqslant 1 - \alpha \right\},$$

where $Z \sim P_0$. Let also $e_{j \cdot M} \in \mathbb{R}^{|M|}$ be the canonical vector with entry 1 corresponding to feature $j \in M$ and $\hat{\sigma}_{j\cdot M} = \sqrt{e_{j\cdot M}^\top (X_M^\top X_M)^{-1} e_{j \cdot M}}$.

Corollary 4.4.1. *Fix $\alpha \in (0,1)$ and $\nu \in (0, \alpha)$. Let*

$$\widehat{\mathcal{M}}_\nu^+ = \left\{ \hat{M}(y') : \left\| X^\top y - X^\top y' \right\|_\infty \leqslant 2 q^\nu \left(\{X_j\}_{j=1}^d \right) \right\}$$

and

$$\widehat{\mathcal{V}}_\nu^+ = \left\{ \frac{e_{j\cdot M}^\top X_M^+}{\hat{\sigma}_{j\cdot M}} : M \in \widehat{\mathcal{M}}_\nu^+, j \in M \right\}.$$

Then,

$$P_\mu \left\{ \theta_{j \cdot \hat{M}} \in \left(\hat{\theta}_{j\cdot \hat{M}} \pm q^{(\alpha - \nu)} \left(\widehat{\mathcal{V}}_\nu^+ \right) \hat{\sigma}_{j \cdot M} \right), \ \forall j \in \hat{M} \right\} \geqslant 1 - \alpha.$$

Therefore, unlike the PoSI method [12], we take a simultaneous correction only over models that seem plausible in hindsight. The correction of Corollary 4.4.1 is at least as tight as the PoSI correction at error level $\alpha - \nu$; note that we can in principle obtain a correction that is at least as tight as the PoSI correction at error level α, by invoking the refined analysis of Theorem 4.4.2. However, this refined correction would require computing both the full PoSI correction and the local correction $q^{(\alpha-\nu)} \left(\widehat{\mathcal{V}}_\nu^+ \right)$, which can be far more computationally demanding than computing only the local correction. As a result, we feel that the correction of Corollary 4.4.1 is more practical.

As a warmup, we instantiate Corollary 4.4.1 for marginal screening, which admits a simple, explicit characterization of $\widehat{\mathcal{M}}_\nu^+$. Then we study selection via the LASSO.

Example 4.4.1 (Marginal screening)**.** *Marginal screening is a simple feature selection method that selects $\hat{M} = \{\hat{i}_1, \dots, \hat{i}_k\}$, where $\hat{i}_j$ is the j-th largest inner product $|X_i^\top y|$, for a pre-specified $k \in [d]$.*

Let $c_{(j)}$ denote the j-th largest inner product $|X_i^\top y|$. Then, it is not difficult to see that $\widehat{\mathcal{M}}_\nu^+$ consists of all subsets of size k of the set

$$\left\{ i \in [d] : |X_i^\top y| \geqslant c_{(k)} - 4 q^\nu (\{X_j\}_{j=1}^d) \right\}.$$

In words, all variables with inner product $|X_i^\top y|$ within a $4q^\nu(\{X_j\}_{j=1}^d)$ margin of $c_{(k)}$ have a plausible chance of being selected. As a result, taking a simultaneous correction over them suffices to get valid inference, while all other variables can be searched through "for free."

Model selection via the LASSO

We discuss a method for locally simultaneous inference after model selection via the LASSO. While we focus on the LASSO, the method can be applied to any selection procedure where the selection event admits a polyhedral representation, such as forward stepwise [164]. We will elaborate on this point later in the section.

Recall that the LASSO solves the following penalized regression problem:

$$\hat{\beta}(y) = \arg\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1,$$

and selects $\hat{M} = \{i \in [d] : \hat{\beta}(y)_i \neq 0\}$. We will write $\hat{\beta}(y) \equiv \hat{\beta}$ when the argument is clear from the context.

The key step in applying Corollary 4.4.1 is to find the set of plausible models $\widehat{\mathcal{M}}_\nu^+$. More precisely, denoting $\mathcal{B}_\nu^\infty = \{y' : \|X^\top y - X^\top y'\|_\infty \leqslant 2q^\nu(\{X_j\}_{j=1}^d)\}$ the relevant neighboring outcome vectors, the set of plausible models is $\widehat{\mathcal{M}}_\nu^+ = \{\hat{M}(y') : y' \in \mathcal{B}_\nu^\infty\}$. To simplify notation we will denote by $s_\nu = 2q^\nu(\{X_j\}_{j=1}^d)$ the radius of $\mathcal{B}_\nu^\infty$.

To find all possible models in $\mathcal{B}_\nu^\infty$, we apply the polyhedral characterization of the LASSO selection event due to Lee et al. [106]. Denoting by $\hat{s} = \text{sign}\left(\hat{\beta}_{\hat{M}}\right)$ the signs of the selected variables in the LASSO solution, Lee et al. show that

$$\{\hat{M} = M, \hat{s} = s\} = \left\{ \begin{pmatrix} A_0^+(M,s) \\ A_0^-(M,s) \\ A_1(M,s) \end{pmatrix} y < \begin{pmatrix} b_0^+(M,s) \\ b_0^-(M,s) \\ b_1(M,s) \end{pmatrix} \right\},$$

for any fixed model-sign pair (M, s), where

$$\begin{aligned} A_0^+(M,s) &= \frac{1}{\lambda} X_{M^c}^\top (I - \Pi_M), \quad b_0^+(M,s) = \mathbf{1} - X_{M^c}^\top (X_M^\top)^+ s; \\ A_0^-(M,s) &= -\frac{1}{\lambda} X_{M^c}^\top (I - \Pi_M), \quad b_0^-(M,s) = \mathbf{1} + X_{M^c}^\top (X_M^\top)^+ s; \\ A_1(M,s) &= -\text{diag}(s)(X_M^\top X_M)^{-1} X_M^\top, \quad b_1(M,s) = -\lambda \text{diag}(s)(X_M^\top X_M)^{-1} s. \end{aligned}$$

Here, $\Pi_M := X_M(X_M^\top X_M)^{-1}X_M^\top$. We will denote the polyhedron above by $P(M, s)$.

At a high level, our approach to finding $\widehat{\mathcal{M}}_\nu^+$ is the following. The Lee et al. characterization shows that the set of outcome vectors for which a model M and sign vector s are realized is a polyhedron. Moreover, for each active constraint of the polyhedron, meaning that the constraint is not redundant in defining the polyhedron, we know exactly which model-sign pair is on the other side of the face (depending on the constraint corresponding to the active face). The basic idea of our procedure is to compute the model-sign pair (and the corresponding polyhedron) at the data y, and then recursively move to neighboring polyhedra until the whole box $\mathcal{B}_\nu^\infty$ is tiled by the visited polyhedra. The set $\widehat{\mathcal{M}}_\nu^+$ is then simply all the models recorded in the visited polyhedra.

The described principle is agnostic to the fact that the polyhedron characterizes the LASSO selection event specifically. In particular, it works for *any* selection procedure that admits a polyhedral representation. Just like in the case of the LASSO, the goal is to enumerate all polyhedra contained in $\mathcal{B}_\nu^\infty$, which encode the different plausible selection events, and this is precisely what our method accomplishes.

In what follows we discuss rules for determining the set of neighboring model-sign polyhedra given the current model-sign polyhedron, which make the core of our procedure. We rely on two types of rules: *exact* screening rules and *safe* screening rules. Exact screening rules are necessary and sufficient to screen out "irrelevant" variables, i.e. those whose inclusion/exclusion does not change when going from the current model-sign region to any neighboring model-sign region: they either remain in the model with the same sign in all neighboring polyhedra or they never enter the model. Safe screening rules are not exact but provide sufficient conditions for screening; we combine them with exact rules to improve computational efficiency. Our safe rules resemble prior work on variable elimination for the LASSO [67, 163], but are fundamentally different as they rely on properties of $\mathcal{B}_\nu^\infty$. It is worth mentioning that the safe rules are LASSO-specific; the exact rules work for general selection strategies with a polyhedral characterization.

We use $\mathcal{B}(M, s)$ to denote the set of model-sign pairs whose corresponding polyhedra neighbor, i.e. share a face with, $P(M, s)$.

Exact screening rules. Exact screening rules proceed by checking for each variable $i \in [d]$ if it can change its inclusion/exclusion status when going from the current model-sign polyhedron $P(M, s)$ to any neighboring polyhedron. In other words, for each variable $i \in M$, they check if there exists a pair $(M', s') \in \mathcal{B}(M, s)$ such that $i \notin M'$; similarly, for each $i \in M^c$, they check if there exists a pair $(M', s') \in \mathcal{B}(M, s)$ such that $i \in M'$, and they additionally identify the corresponding sign of variable i if such a pair exists.

The core idea of exact screening rules is to find the minimal representation of $P(M, s) \cap \mathcal{B}_\nu^\infty$. That is, the goal is to prune all redundant constraints coming from $P(M, s)$; the inequalities that remain are "active" and indicate that the variables corresponding to those constraints can enter or leave the model in one of the neighboring polyhedra. In Algorithm 9 we use a standard solution to finding a minimal polyhedral representation, which relies on solving one linear program for each constraint whose redundancy is being checked.

Safe screening rules. Safe rules serve to speed up the search for a minimal representation of a polyhedron corresponding to a model-sign pair.

For all $y' \in P(M, s)$, the LASSO optimality conditions imply that the LASSO solution is locally linear, namely

$$\hat{\beta}(y') = \beta_{(M,s)}(y') := (X_M^\top X_M)^{-1}(X_M^\top y' - \lambda s).$$

Note that, while $\beta_{(M,s)}(y')$ is equal to the LASSO solution for $y' \in P(M, s)$, it can be computed for $y' \notin P(M, s)$. We use this characterization to design the safe screening rules.

Algorithm 7 Locally simultaneous inference for the LASSO

```
input: design matrix $X$, outcome vector $y$, penalty $\lambda$, error level $\alpha$, parameter $\nu \in (0, \alpha)$
output: set of plausible models $\widehat{\mathcal{M}}_\nu^+$
  Compute width of $\mathcal{B}_\nu^\infty$: $s_\nu = 2q^\nu(\{X_j\}_{j=1}^d)$
  Compute LASSO solution: $\hat{\beta} = \arg\min_\beta \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$
  Let $\hat{M} = \mathrm{supp}(\hat{\beta})$, $\hat{s} = \mathrm{sign}(\hat{\beta}_{\hat{M}})$
  Initialize $\mathcal{P}_{\mathrm{todo}} \leftarrow \{(\hat{M}, \hat{s})\}, \mathcal{P}_\nu^+ \leftarrow \emptyset$
  while $\mathcal{P}_{\mathrm{todo}} \neq \emptyset$ do
    Take any pair $(M, s) \in \mathcal{P}_{\mathrm{todo}}$
    Update $(M, s)$ as visited: $\mathcal{P}_{\mathrm{todo}} \leftarrow \mathcal{P}_{\mathrm{todo}} \setminus \{(M, s)\}$, $\mathcal{P}_\nu^+ \leftarrow \mathcal{P}_\nu^+ \cup \{(M, s)\}$
    $\mathcal{I}_{\mathrm{safe}}(M, s) \leftarrow \mathrm{SafeScreening}(X, y, (M, s))$    (Alg. 8)
    $\mathcal{B}(M, s) \leftarrow \mathrm{ExactScreening}(X, y, (M, s), \mathcal{I}_{\mathrm{safe}}(M, s))$    (Alg. 9)
    $\mathcal{P}_{\mathrm{todo}} \leftarrow \mathcal{P}_{\mathrm{todo}} \cup (\mathcal{B}(M, s) \setminus \mathcal{P}_\nu^+)$
  end while
  Return $\widehat{\mathcal{M}}_\nu^+ = \{M : \exists s \text{ s.t. } (M, s) \in \mathcal{P}_\nu^+\}$
```

Lemma 4.4.2 (Safe exclusion)**.** *Fix a model-sign pair (M, s). Let*

$$\mathcal{I}_{\mathrm{safe}}^-(M, s) := \left\{j \in M^c : |X_j^\top(y - X_M\beta_{(M,s)}(y))| < \lambda - s_\nu\left(1 + \|X_j^\top X_M(X_M^\top X_M)^{-1}\|_1\right)\right\}.$$

Then, for any $j \in \mathcal{I}_{\mathrm{safe}}^-(M, s)$, variable j cannot enter the model in any of the neighboring polyhedra:

$$\forall (M', s') \in \mathcal{B}(M, s),\ j \notin M'.$$

Lemma 4.4.3 (Safe inclusion)**.** *Fix a model-sign pair (M, s). Let*

$$\mathcal{I}_{\mathrm{safe}}^+(M, s) = \left\{j \in M : |\beta_{j\cdot(M,s)}(y)| > s_\nu \left\|e_{j\cdot(M,s)}^\top(X_M^\top X_M)^{-1}\right\|_1\right\}.$$

Then, for any $j \in \mathcal{I}_{\mathrm{safe}}^+(M, s)$, variable j cannot exit the model in any of the neighboring polyhedra:

$$\forall (M', s') \in \mathcal{B}(M, s),\ j \in M'.$$

Lemma 4.4.2 and Lemma 4.4.3 show that the safe screening subroutine is valid. Putting everything together, we formalize the guarantees of Algorithm 7 in Theorem 4.4.5.

Theorem 4.4.5. *Algorithm 7 returns exactly the set of plausible models, i.e.*

$$\widehat{\mathcal{M}}_\nu^+ = \left\{\hat{M}(y') : \|X^\top y - X^\top y'\|_\infty \leqslant 2q^\nu(\{X_j\}_{j=1}^d)\right\}.$$

Putting together Theorem 4.4.5 and Corollary 4.4.1, we conclude that it suffices to take a simultaneous correction in the sense of Berk et al. [12] at error level $\alpha - \nu$, *only* over the local model set $\widehat{\mathcal{M}}_\nu^+$, to get a valid confidence region for $\theta_{\hat{M}}$.

To state the subroutines of Algorithm 7 for safe and exact screening, we introduce the necessary notation. We denote by $A_0^{+j}(M,s)$ the row in $A_0^{+}(M,s)$ corresponding to the variable $j \in M^c$. We adopt analogous definitions for $A_0^{-j}(M,s)$ and $A_1^{j}(M,s)$ (where in the latter case we consider $j \in M$). For $j \in M^c$, we will use $P^{\backslash\{+j\}}(M,s)$ (resp. $P^{\backslash\{-j\}}(M,s)$) to denote the polyhedron $P(M,s)$ with constraint $(A_0^{+j}(M,s), b_0^{+j}(M,s))$ (resp. $(A_0^{-j}(M,s), b_0^{-j}(M,s))$) removed. We similarly use $P^{\backslash\{j\}}(M,s)$ for $j \in M$. We use $(M,s)^{-j}$ to denote the model-sign pair obtained by removing variable $j \in M$ and the corresponding sign. Similarly, we use $(M,s)^{+(j,+1)}$ to denote the model-sign pair obtained by adding variable $j \in M^c$ to M with a positive sign. We use $(M,s)^{+(j,-1)}$ analogously, only the corresponding sign is negative.

Algorithm 8 SafeScreening

```
input: design matrix $X$, outcome vector $y$, current model-sign pair $(M,s)$
output: safely screened variables
  $\mathcal{I}_{\text{safe}}(M,s)$
  Compute extrapolated solution at $y$, $\beta_{(M,s)}(y) = (X_M^\top X_M)^{-1}(X_M^\top y - \lambda s)$
  $\mathcal{I}^{+}_{\text{safe}}(M,s) \leftarrow \{j \in M : |\beta_{j\cdot(M,s)}(y)| > s_\nu \|e_{j\cdot(M,s)}^\top (X_M^\top X_M)^{-1}\|_1\}$
  $\mathcal{I}^{-}_{\text{safe}}(M,s) \leftarrow \{j \in M^c : |X_j^\top (y - X_M \beta_{(M,s)}(y))| < \lambda - s_\nu(1 + \|X_j^\top X_M (X_M^\top X_M)^{-1}\|_1)\}$
  Return $\mathcal{I}_{\text{safe}}(M,s) \leftarrow \mathcal{I}^{+}_{\text{safe}}(M,s) \cup \mathcal{I}^{-}_{\text{safe}}(M,s)$
```

Algorithm 9 ExactScreening

```
input: design matrix $X$, outcome vector $y$, current model-sign pair $(M,s)$, (optionally)
safely screened variables $\mathcal{I}_{\text{safe}}(M,s)$
output: neighboring model-sign pairs $\mathcal{B}(M,s)$
  Initialize $\mathcal{B}(M,s) \leftarrow \emptyset$
  $\forall j \in M^c \setminus \mathcal{I}_{\text{safe}}(M,s)$, compute constraint $(A_0^{+j}(M,s), b_0^{+j}(M,s)), (A_0^{-j}(M,s), b_0^{-j}(M,s))$
  $\forall j \in M \setminus \mathcal{I}_{\text{safe}}(M,s)$, compute constraint $(A_1^{j}(M,s), b_1^{j}(M,s))$
  for $j \in [d] \setminus \mathcal{I}_{\text{safe}}(M,s)$ do
    if $j \in M$ then
      Solve LP: $\texttt{Val} = \max_z z^\top A_1^{j}(M,s)$ s.t. $z \in P^{\backslash\{j\}}(M,s)$
      If $\texttt{Val} > b_1^{j}(M,s)$, add $(M,s)^{-j}$ to $\mathcal{B}(M,s)$
    else if $j \in M^c$ then
      Solve LP: $\texttt{Val} = \max_z z^\top A_0^{+j}(M,s)$ s.t. $z \in P^{\backslash\{+j\}}(M,s)$
      If $\texttt{Val} > b_0^{+j}(M,s)$, add $(M,s)^{+(j,+1)}$ to $\mathcal{B}(M,s)$
      Solve LP: $\texttt{Val} = \max_z z^\top A_0^{-j}(M,s)$ s.t. $z \in P^{\backslash\{-j\}}(M,s)$
      If $\texttt{Val} > b_0^{-j}(M,s)$, add $(M,s)^{+(j,-1)}$ to $\mathcal{B}(M,s)$
    end if
  end for
  Return $\mathcal{B}(M,s)$
```

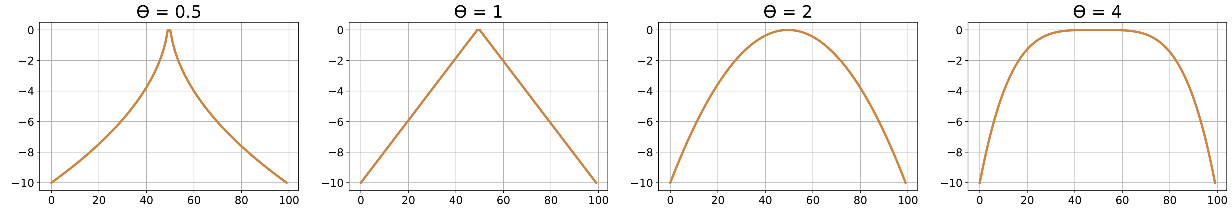


Figure 4.10: Mean outcome μ for different problem parameters θ, at scale level $C = 10$.

4.4.4 Numerical evaluation

We compare locally simultaneous inference to standard simultaneous inference, the conditional method due to Lee et al. [106], and the hybrid refinement of conditional inference due to Andrews et al. [2]. Throughout we apply the version of locally simultaneous inference from Theorem 4.4.1 with $\nu = 0.1\alpha$. In all figures comparing interval widths we plot the median width over 100 trials, together with the 5% and 95% quantile, plotted as error bars around the median. The target error level is $\alpha = 0.1$ throughout.

Inference on the winner

We begin by studying the problem of inference on the winner from Section 4.4.2. We generate the mean outcome μ as a smooth curve; this simulates a setting where nearby entries of μ are similar, such as when the data is a time series or when neighboring entries of μ correspond to outcomes in neighboring subgroups (e.g., neighboring age groups). We vary the shape of the mean outcome vector μ, thereby making inference more or less challenging for the different methods. We let $\mu_i \propto -|i - 0.5(m+1)|^\theta$ for $i \in [m]$, where $\theta > 0$ varies the sharpness of μ. Small θ corresponds to the case where the winner stands out, while large θ makes the mean outcome flat, implying that many observations have a plausible chance of being selected as the winner. The other tuning parameter is $C > 0$: we rescale the mean μ so that the difference between the minimum and maximum entry of μ is equal to C. When C is large, μ gets "stretched out" and, as a result, there are fewer candidates that can plausibly be selected. In Figure 4.10 we plot the shape of μ for different values of θ, at $C = 10$.

Parametric case. In the first setting, we generate the vector of observations as $y = \mu + \xi$, where $\xi \sim \mathcal{N}(0, I_m)$. In Figure 4.11, we plot the interval width resulting from locally simultaneous, simultaneous, conditional, and hybrid inference for varying $\theta \in \{0.5, 1, 2, 4\}$, $C \in \{10, 30, 50, 70\}$, and $m \in \{10, 10^2, 10^3, 10^4\}$. The mean μ has range C at $m = 10$ and for higher m it is not renormalized to range C; the purpose of increasing m is to demonstrate the behavior of the different methods when the number of irrelevant observations (i.e., those far from the winning observation) increases. We observe that conditional inference exhibits high variability for all problem parameters, and as θ grows—meaning μ becomes flat—the median

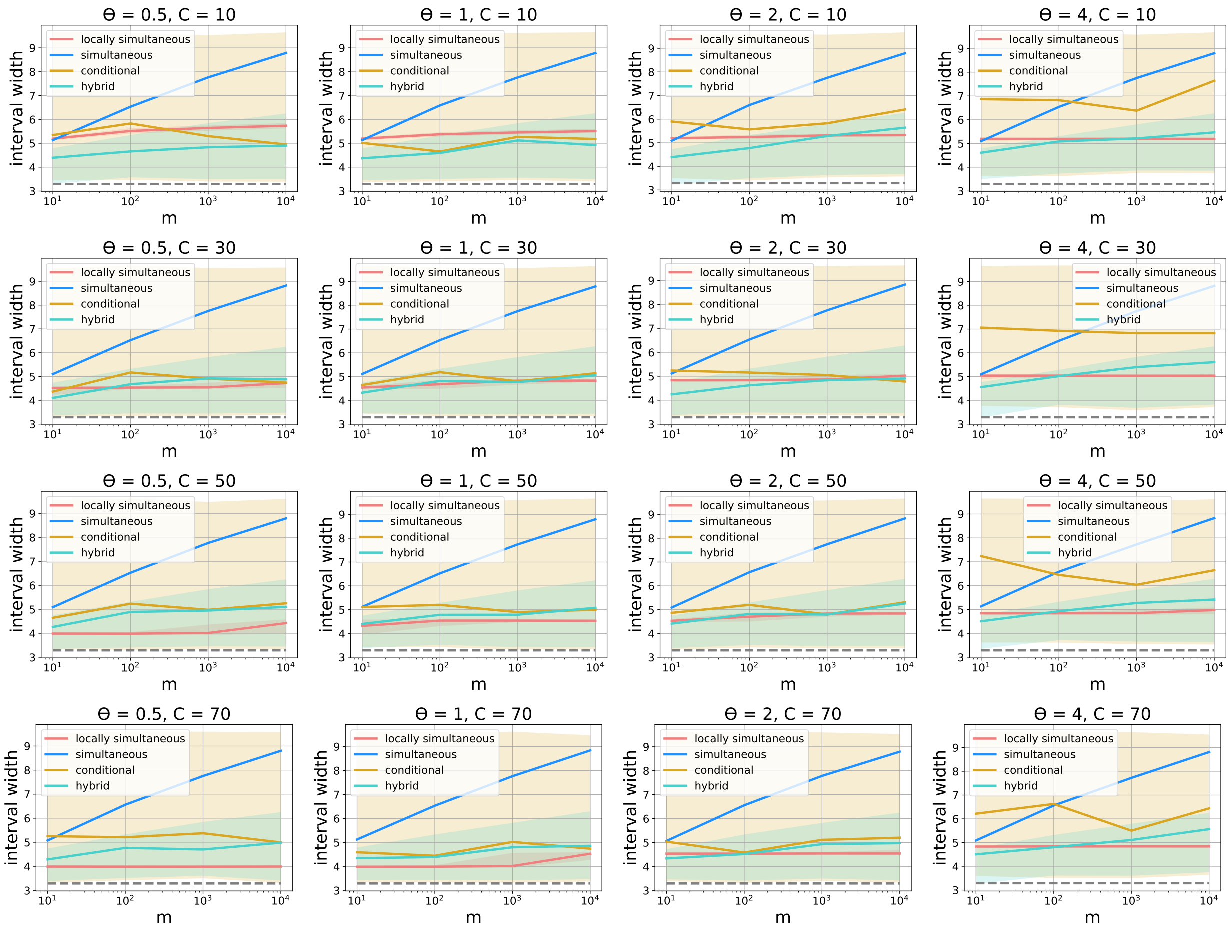


Figure 4.11: Interval width achieved by locally simultaneous, fully simultaneous, conditional, and hybrid inference in the problem of inference on the winner. The dashed line corresponds to nominal interval widths.

intervals become large. Simultaneous inference is by construction only sensitive to changes in m, and its intervals grow with m despite the fact that only the number of irrelevant observations grows. Locally simultaneous inference is most sensitive to changes in C: as μ is stretched over a larger range, the method finds fewer plausible candidates and thus leads to smaller intervals. Moreover, it is virtually insensitive to increasing m. The hybrid approach exhibits high variability like the conditional approach (albeit to a more moderate extent) and its intervals grow with m because, as $m \to \infty$, the hybrid method reduces to standard conditional inference.

Nonparametric case. We emphasized that locally simultaneous inference is rigorously applicable in nonparametric settings, while conditional approaches are not. Still, it might

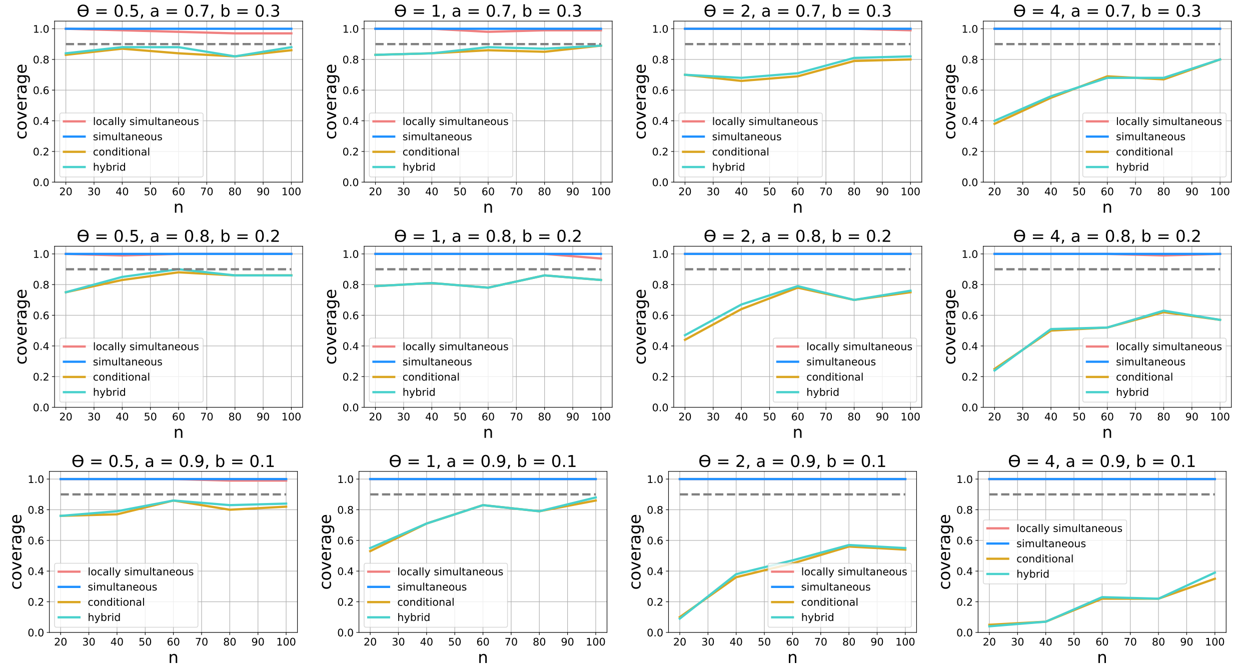


Figure 4.12: Coverage of locally simultaneous, fully simultaneous, conditional, and hybrid inference when the noise is sampled from Beta(a, b). The conditional and hybrid approaches use a normal approximation; the locally simultaneous and fully simultaneous approaches use nonparametric, finite-sample-valid confidence intervals due to Waudby-Smith and Ramdas [177]. The target coverage is 0.9, indicated by the dashed line.

seem like a reasonable heuristic to apply conditional inference after a normal approximation based on the CLT. We test this heuristic empirically, comparing to a nonparametric application of locally and fully simultaneous inference. We observe that the heuristic application of conditional methods can severely undercover the target.

We fix $C = 20$, $m = 100$, and vary θ to obtain the mean vector μ. Given μ, we generate n i.i.d. samples $y^{(1)}, \dots, y^{(n)}$, where $y^{(j)} = \mu + \xi^{(j)}$ and $\xi^{(j)}$ has i.i.d. entries sampled from Beta(a, b). To apply the locally and fully simultaneous methods, we use the betting-based confidence intervals by Waudby-Smith and Ramdas [177] (Theorem 3), together with a Bonferroni correction. To form the acceptance region of the locally simultaneous method, we use the Bentkus concentration inequality [11]. In Figure 4.12 we plot the coverage of all four approaches for varying a, b, and sample size n. We observe that, as θ grows, the conditional methods have diminishing coverage. This confirms the need for a more robust, nonparametrically applicable correction. In contrast, the two simultaneous methods have valid coverage and typically overcover, which is to be expected given the use of nonparametric concentration inequalities. In Figure 4.13 we plot the interval width implied by the four

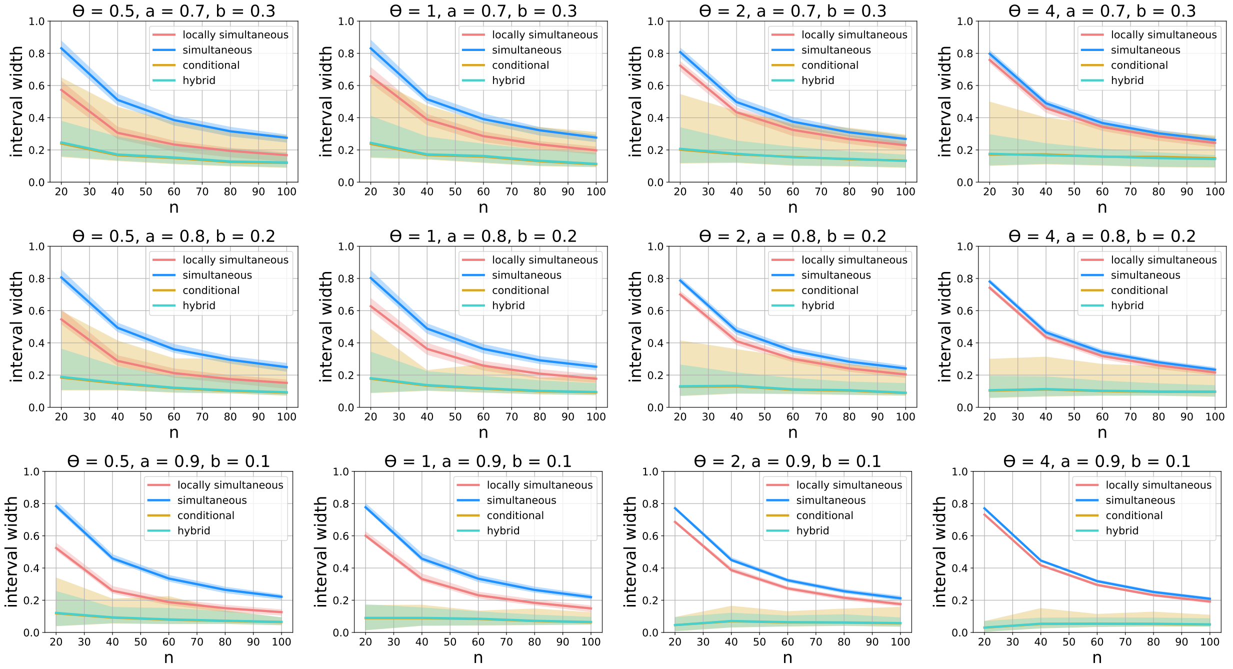


Figure 4.13: Interval width achieved by locally simultaneous, fully simultaneous, conditional, and hybrid inference when the noise is sampled from Beta(a, b). The conditional and hybrid approaches use a normal approximation; the locally simultaneous and fully simultaneous approaches use nonparametric, finite-sample-valid confidence intervals due to Waudby-Smith and Ramdas [177].

methods. The conditional methods yield much smaller intervals, but this comes at the cost of invalid coverage, as shown in Figure 4.12. The locally simultaneous intervals are consistently smaller than the fully simultaneous intervals, with the improvement being more pronounced when there are few plausible candidates, that is, when θ is small. Moreover, as n grows, the locally simultaneous intervals gradually approach the conditional intervals; this makes sense seeing that the coverage of the conditional methods improves with n.

File-drawer problem

The next problem we consider is the file-drawer problem from Section 4.4.2. As alluded to earlier, the conditional and hybrid approaches provide inference for one real-valued parameter at a time and it is unclear how to generalize them to multi-dimensional problems without resorting to a Bonferroni correction. In contrast, locally simultaneous inference is able to adapt to the dependencies in the data.

To demonstrate this, we consider $y = \mu + \xi$, where $\xi \sim \mathcal{N}(0, \Sigma)$ is a Gaussian noise

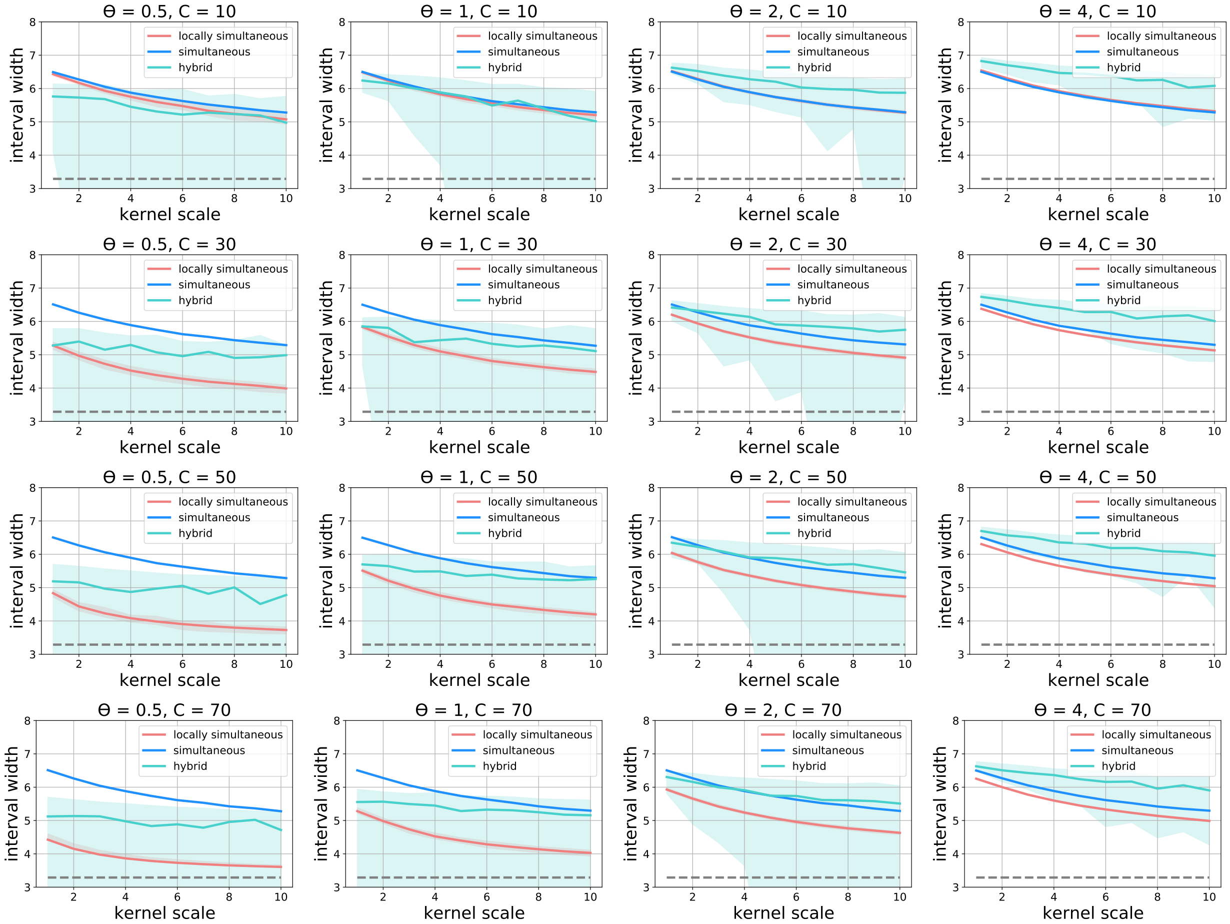


Figure 4.14: Interval width achieved by locally simultaneous, fully simultaneous, and hybrid inference in the file-drawer problem. Conditional inference achieves much wider intervals and is thus not included in the plots. The dashed line corresponds to nominal interval widths.

process with the RBF kernel, $\Sigma_{ij} = \exp\left(-\frac{|i-j|^2}{2\phi^2}\right)$; ϕ is the key parameter that we vary. As ϕ gets larger, the errors become more dependent. We generate μ as in the first problem setting, again varying θ and C.

First, we observe that the conditional approach is exceptionally fragile in this problem setting: its intervals are consistently much larger than the intervals of the other competitors, often even of a different order of magnitude. For this reason, we omit the conditional approach from the comparison. In Figure 4.14 we plot the interval widths of locally simultaneous, simultaneous, and hybrid inference. We set $T = -1$ and vary the kernel scale ϕ, as well as θ and C, which control the shape of μ. We combine the hybrid method with a Bonferroni correction over the selected set. We observe that the simultaneous and locally simultaneous methods are indeed able to adapt to the kernel scale. Moreover, as in the previous problem setting, increasing θ makes the problem more challenging for the hybrid method, and as C increases the problem becomes easier for locally simultaneous inference.

Inference after selection via the LASSO

Next, we look at the problem of inference after model selection via the LASSO.

Already when the dimension d is greater than 20, the number of models admissible for selection exceeds 10^6, making the fully simultaneous PoSI method of Berk et al. [12] prohibitively computationally expensive. Here we show that the set of plausible models $\widehat{\mathcal{M}}_\nu^+$ can be much smaller than the set of all subsets of $[d]$ when the true data-generating model is sparse, making locally simultaneous inference both powerful and computationally tractable.

In Figure 4.15, we consider the following data-generating process. We generate the design matrix to have i.i.d. standard normal entries and normalize the columns to have norm 1. We let $y = X\beta + \xi$, where $\xi \sim \mathcal{N}(0, I_n)$ and β has $\lceil s \cdot d \rceil$ nonzero entries, where we vary the sparsity parameter s. Of the $\lceil s \cdot d \rceil$ nonzero entries, we take half of them to be "weak", specifically equal to λ, and half of them to be "strong", specifically equal to 2λ. We let λ have the usual scaling of $\sim \sqrt{2\log(e \cdot d)}$. In particular, we fix $\lambda = 6\sqrt{2\log(e \cdot d)}$ and $n = 1000$. In this parameter regime, we observe that the plausible models are typically those models that always include the strong variables, never include the irrelevant variables, and contain an arbitrary subset of the weak variables. We only compare locally simultaneous inference to conditional inference, seeing that fully simultaneous inference is computationally challenging for the values of d we consider. As before, we observe that the conditional approach exhibits high variability. Moreover, the median interval width implied by the locally simultaneous approach is noticeably smaller.

That being said, the locally simultaneous solution for the LASSO has computational disadvantages. Its complexity scales with the number of plausible model-sign pairs and there can be up to 3^d such corresponding pairs, which means that in the worst case the search for all plausible models can be fairly slow. A reasonable remedy is to introduce a parameter $P_{\max}$ such that, if the size of $\mathbb{P}_{\text{todo}}$ in Algorithm 7 exceeds $P_{\max}$, the search for new model-sign pairs stops and the procedure simply runs the PoSI method of Berk et al. at error level $\alpha - \nu$. We implement this strategy in Figure 4.16. Specifically, we let $d = 10$ and generate X and y as before, only now $\beta_i = \varepsilon$ for $i \in \{1, \ldots, 5\}$ and $\beta_i = 0$ for $i \in \{6, \ldots, 10\}$. We set $\lambda = \lambda_0\sqrt{2\log(e \cdot d)}$ and vary λ_0 and ε/λ_0. When $\varepsilon/\lambda_0 = 1$, the non-nulls of β are approximately at the threshold of being selected, and as ε/λ grows the true model becomes more obvious. To speed up locally simultaneous inference, we set $P_{\max} = 2000$. As in the experiments on inference on the winner, we observe that locally simultaneous inference is preferred when the data is near the selection boundary, which happens when λ_0 or ε/λ is small. As the selection becomes more obvious, conditional inference becomes more powerful.

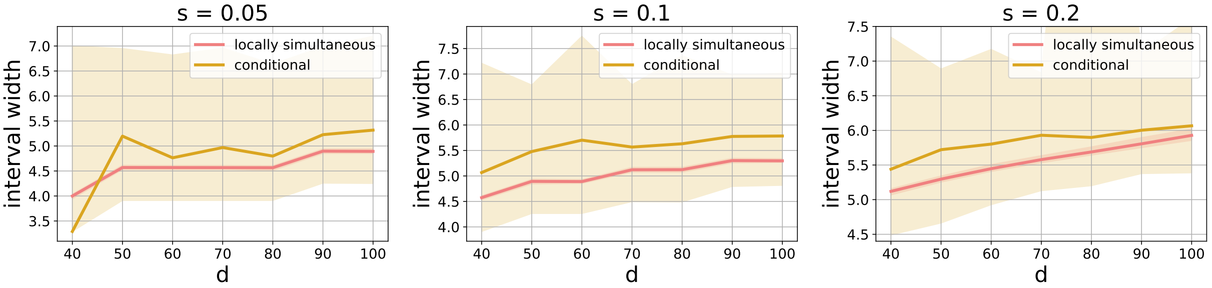


Figure 4.15: Interval width achieved by locally simultaneous and conditional inference in the problem of inference after selection via the LASSO, when the true underlying signal is s-sparse.

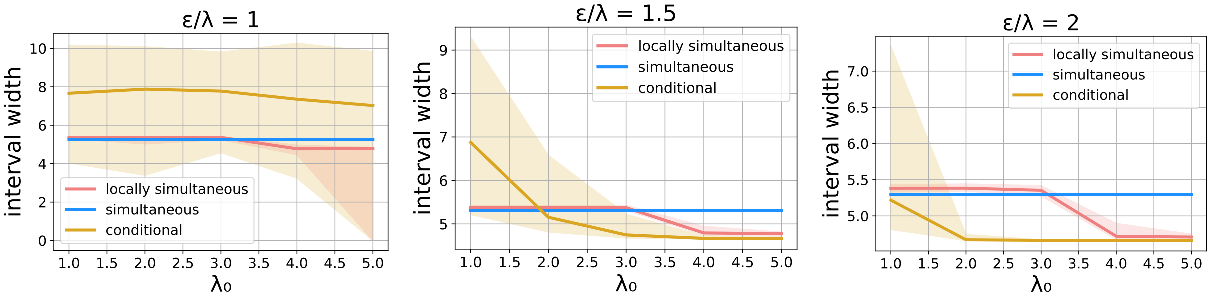


Figure 4.16: Interval width achieved by locally simultaneous, fully simultaneous, and conditional inference in the problem of inference after selection via the LASSO, when we vary to ratio of signal strength to regularization.

Experiments on real climate data

Finally, we conduct experiments on a real climate dataset [143]. The dataset contains hourly measurements of temperature from 1999 to 2018 across a discrete grid of locations on Earth. The grid is obtained by pairing 32 latitude coordinates with 64 longitude coordinates. We model the measurements from the 20 years as i.i.d. draws from an underlying distribution. We again compare locally simultaneous, fully simultaneous, conditional, and hybrid inference. To be able to apply the conditional and hybrid approaches, we model the draws as i.i.d. multivariate Gaussians. We use older data, from 1979 to 1998, to estimate the Gaussian covariance. We study two types of selection: based on time and based on location.

In the first set of experiments we compute the average temperature on Earth (averaged over all locations on the grid) and look at the resulting time series. For each year we take one measurement per day, evaluated at noon, resulting in a series of 365 entries. We ask for inference on the warmest day, coldest day, and all days with temperature above 8. In

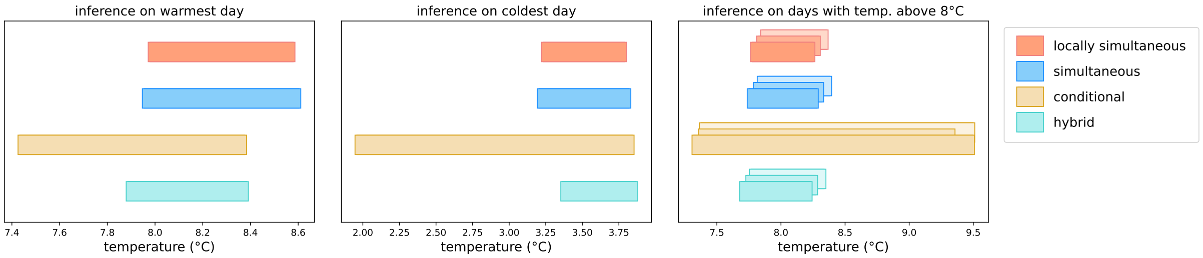


Figure 4.17: Intervals for the mean temperature constructed via locally simultaneous, fully simultaneous, conditional, and hybrid inference, for selections based on time.

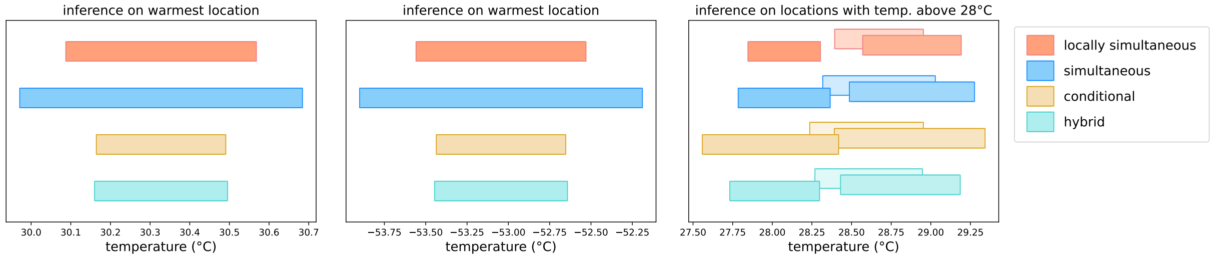


Figure 4.18: Intervals for the mean temperature constructed via locally simultaneous, fully simultaneous, conditional, and hybrid inference, for selections based on location.

the second set of experiments we compute the average annual temperature and look at its distribution over the different recorded locations on Earth. Similarly to the first set of experiments, we ask for inference on the warmest location, coldest location, and all locations with temperature above 28. We plot the resulting intervals in Figure 4.17 and Figure 4.18, respectively. For the two file-drawer problems, for interpretability we only visualize three intervals, corresponding to the first three days/locations where the temperature exceeds the critical threshold. In the first set of experiments, we observe that conditional inference leads to significantly wider intervals than any of the three alternatives; in contrast, in the first two problems of the second set of experiments the winner stands out and hence the conditional method outperforms the other approaches. The locally simultaneous approach gives narrower intervals than the fully simultaneous approach, as expected. The hybrid approach leads to smaller intervals than the locally simultaneous approach in some settings and wider in others. Critically, however, the hybrid approach is only applicable because we imposed an assumption of Gaussianity, while locally simultaneous inference would be applicable even nonparametrically.

4.5 Deferred proofs

4.5.1 Auxiliary lemmas

Lemma 4.5.1 (Composition theorem [23, 56])**.** *Let $\mathcal{A}^{(k)}$ be the adaptively composed algorithm after k rounds (Alg. 4). Fix two vectors $y, y' \in \mathbb{R}^n$ and suppose that $\mathcal{A}_t(a_1, \dots, a_{t-1}, y) \approx_{\eta,\tau} \mathcal{A}_t(a_1, \dots, a_{t-1}, y')$, for every fixed sequence $a_1, \dots, a_{t-1}$, and all $t \in [k]$. Then,*

(a) $\mathcal{A}^{(k)}(y) \approx_{k\eta,k\tau} \mathcal{A}^{(k)}(y')$,

(b) $\mathcal{A}^{(k)}(y) \approx_{\frac{1}{2}k\eta^2+\sqrt{2k\log(1/\delta)}\eta,k\tau+\delta} \mathcal{A}^{(k)}(y')$, for all $\delta \in (0,1)$.

4.5.2 Proof of Lemma 4.3.1

Denote by $\hat{\Gamma}_0$ the oracle from Definition 4.3.2 and let $E = \{\omega \in \mathbb{R}^n : \hat{\Gamma}(\omega) \approx_{\eta,\tau} \hat{\Gamma}_0\}$. Fix an event $\mathcal{O} \subseteq \mathbb{R}^n \times \mathcal{S}$, and let $\mathcal{O}_\omega = \{\Gamma \in \mathcal{S} : (\omega, \Gamma) \in \mathcal{O}\}$. Notice that $\mathbf{1}\{(y, \hat{\Gamma}(y)) \in \mathcal{O}\} = \mathbf{1}\{\hat{\Gamma}(y) \in \mathcal{O}_y\}$, and hence $\mathbb{E}[\mathbf{1}\{(y, \hat{\Gamma}(y)) \in \mathcal{O}\}|y] = \mathbb{E}[\mathbf{1}\{\hat{\Gamma}(y) \in \mathcal{O}_y\}|y]$.

With this, we can write:

$$\begin{aligned} P\{(y, \hat{\Gamma}(y)) \in \mathcal{O}, y \in E\} &= \mathbb{E}[\mathbb{E}[\mathbf{1}\{\hat{\Gamma}(y) \in \mathcal{O}_y\}|y]\mathbf{1}\{y \in E\}] \\ &= \mathbb{E}[P\{\hat{\Gamma}(y) \in \mathcal{O}_y|y\}\mathbf{1}\{y \in E\}] \\ &\leqslant \mathbb{E}[(e^\eta P\{\hat{\Gamma}_0 \in \mathcal{O}_y|y\} + \tau)\mathbf{1}\{y \in E\}] \\ &= \mathbb{E}[(e^\eta \mathbf{1}\{\hat{\Gamma}_0 \in \mathcal{O}_y\} + \tau)\mathbf{1}\{y \in E\}] \\ &\leqslant e^\eta P\{(y, \hat{\Gamma}_0) \in \mathcal{O}, y \in E\} + \tau. \end{aligned}$$

Since $P\{y \in E\} \geqslant 1 - \nu$, we can conclude:

$$\begin{aligned} P\{(y, \hat{\Gamma}(y)) \in \mathcal{O}\} &= P\{(y, \hat{\Gamma}(y)) \in \mathcal{O}, y \in E\} + P\{(y, \hat{\Gamma}(y)) \in \mathcal{O}, y \notin E\} \\ &\leqslant P\left\{(y, \hat{\Gamma}(y)) \in \mathcal{O}, y \in E\right\} + \nu \\ &\leqslant e^\eta P\{(y, \hat{\Gamma}_0) \in \mathcal{O}, y \in E\} + \tau + \nu \\ &\leqslant e^\eta P\{(y, \hat{\Gamma}_0) \in \mathcal{O}\} + \tau + \nu. \end{aligned}$$

4.5.3 Proof of Theorem 4.3.1

By Lemma 4.3.1, we know that

$$\begin{aligned} P\{\theta_{\hat{\Gamma}} \notin C_{\hat{\Gamma}}^{\delta e^{-\eta}}\} &\leqslant e^\eta P\{\theta_{\hat{\Gamma}_0} \notin C_{\hat{\Gamma}_0}^{\delta e^{-\eta}}\} + \tau + \nu \\ &= e^\eta\, \mathbb{E}\left[P\left\{\theta_{\hat{S}_0} \notin C_{\hat{\Gamma}_0}^{\delta e^{-\eta}} \;\middle|\; \hat{\Gamma}_0\right\}\right] + \tau + \nu, \end{aligned}$$

where $C^{\delta e^{-\eta}}_{\hat{\Gamma}_0}$ are confidence intervals computed on y and $\hat{\Gamma}_0$ is an oracle selection independent of y. By the construction of $C^{\delta e^{-\eta}}_{\hat{\Gamma}_0}$, we know $P\left\{\theta_{\hat{\Gamma}_0} \notin C^{\delta e^{-\eta}}_{\hat{\Gamma}_0} \,\middle|\, \hat{\Gamma}_0\right\} \leqslant \delta e^{-\eta}$, and therefore

$$P\{\theta_{\hat{\Gamma}} \notin C^{\delta e^{-\eta}}_{\hat{\Gamma}}\} \leqslant e^{\eta} e^{-\eta}\delta + \tau + \nu = \delta + \tau + \nu.$$

4.5.4 Proof of Proposition 4.3.1

Denote by $\mathcal{M}_s$ the set of all models of size at most s and fix any $\tau \in (0,1)$. Let $y' \sim P_y$ be an i.i.d. copy of y. Define the set of bad models to be

$$\mathcal{M}^* = \left\{ M \in \mathcal{M}_s \ : \ \exists \omega^* \in \operatorname{supp}(P_y) \text{ such that } \frac{P\{\hat{M}(\omega^*) = M\}}{P\{\hat{M}(y') = M\}} \geqslant \frac{\sum_{k=1}^{s} \binom{d}{k}}{\tau} \right\}.$$

By definition, we see

$$P\{\hat{M}(y') \in \mathcal{M}^*\} \leqslant \sum_{M \in \mathcal{M}^*} P\{\hat{M}(y') = M\} \leqslant \tau,$$

which follows by taking a union bound over all $\sum_{k=1}^{s} \binom{d}{k}$ possible models. Consequently, for any event $\mathcal{O} \subseteq \mathbb{R}^n \times \mathcal{M}_s$ such that $\{M : \exists \omega \text{ s.t. } (\omega, M) \in \mathcal{O}\} \subseteq \mathcal{M}^*$, we have

$$P\{(y, \hat{M}(y)) \in \mathcal{O}\} \leqslant P\{\hat{M}(y) \in \mathcal{M}^*\} = P\{\hat{M}(y') \in \mathcal{M}^*\} \leqslant \tau.$$

Now denote $\mathcal{O}_\omega = \{M \in \mathcal{M}_s : (\omega, M) \in \mathcal{O}\}$, and notice that $\{(y, \hat{M}(y)) \in \mathcal{O}\} = \{\hat{M}(y) \in \mathcal{O}_y\}$. Then, for all $\mathcal{O} \subseteq \mathbb{R}^n \times \mathcal{M}_s$ such that $\{M : \exists \omega \text{ s.t. } (\omega, M) \in \mathcal{O}\} \cap \mathcal{M}^* = \emptyset$, we know

$$\begin{aligned} P\{(y, \hat{M}(y)) \in \mathcal{O}\} &= P\{\hat{M}(y) \in \mathcal{O}_y\} = \mathbb{E}\left[P\{\hat{M}(y) \in \mathcal{O}_y | y\}\right] \\ &\leqslant \frac{\sum_{k=1}^{s} \binom{d}{k}}{\tau} \mathbb{E}\left[P\{\hat{M}(y') \in \mathcal{O}_y | y\}\right] = \frac{\sum_{k=1}^{s} \binom{d}{k}}{\tau} P\{(y, \hat{M}(y')) \in \mathcal{O}\}. \end{aligned}$$

Finally, take an arbitrary $\mathcal{O} \subseteq \mathbb{R}^n \times \mathcal{M}_s$, and partition it as follows:

$$\mathcal{O}_{\text{bad}} = \{(\omega, M) \in \mathcal{O} : M \in \mathcal{M}^*\}, \quad \mathcal{O}_{\text{good}} = \{(\omega, M) \in \mathcal{O} : M \notin \mathcal{M}^*\}.$$

Putting everything together, we have shown

$$\begin{aligned} P\{(y, \hat{M}(y)) \in \mathcal{O}\} &= P\{(y, \hat{M}(y)) \in \mathcal{O}_{\text{bad}}\} + P\{(y, \hat{M}(y)) \in \mathcal{O}_{\text{good}}\} \\ &\leqslant \tau + \frac{\sum_{k=1}^{s} \binom{d}{k}}{\tau} P\{(y, \hat{M}(y')) \in \mathcal{O}\}. \end{aligned}$$

In other words, we can conclude that $(y, \hat{M}(y)) \approx_{\eta,\tau} (y, \hat{M}(y'))$, with $\eta = \log\left(\frac{\sum_{k=1}^{s} \binom{d}{k}}{\tau}\right) = O(s \log(d/s)) + \log(1/\tau)$, as desired.

Applying the same steps as in Theorem 4.3.1 allows us to conclude that $C_{j \cdot \hat{M}}(K_{\hat{M}, \delta e^{-\eta}}) = \left(\hat{\theta}_{j \cdot \hat{M}} \pm K_{\hat{M}, \delta e^{-\eta}} \hat{\sigma}_{j \cdot \hat{M}}\right)$, where $\eta = O(s \log(d/s)) + \log(1/\tau)$, are valid confidence intervals at level $\delta + \tau$.

A related argument is given in Theorem 6 of Dwork et al. [53].

4.5.5 Proof of Proposition 4.3.2 (LASSO stability)

For the sake of readability, we denote the squared loss, rescaled by $\hat{\sigma}$, by $L(\beta; X, y) := \frac{1}{n\hat{\sigma}}\|y - X\beta\|_2^2$; hence, $\nabla L(\beta; X, y) = \frac{2}{n\hat{\sigma}}X^\top(y - X\beta)$. Also, we denote by $S_{C_1} := C_1 \cdot \{\pm e_i\}_{i=1}^d$ the set of $2d$ extreme points of the ℓ_1-ball in $\mathbb{R}^d$, scaled by the LASSO constraint C_1. Similarly, we let $S_{C_1}^+ := C_1 \cdot \{e_i\}_{i=1}^d$ denote half of the points in S_{C_1} that correspond to the extreme points with non-negative coordinates.

Let $y \sim \mathcal{N}(\mu, \sigma^2 I)$. Fix $t \in [k]$ and β such that $\|\beta\|_1 \leqslant C_1$. For all $\phi \in S_{C_1}$, we have

$$\begin{aligned}\phi^\top(\nabla L(\beta; X, y) - \nabla L(\beta; X, \mu)) &= \phi^\top \left(\frac{2}{n\hat{\sigma}}X^\top(y - X\beta) - \frac{2}{n\hat{\sigma}}X^\top(\mu - X\beta)\right)\\ &= \frac{2}{n\hat{\sigma}}\phi^\top X^\top(y - \mu).\end{aligned}$$

Notice that $\|X\phi\|_2 \leqslant C_1\|X\|_{2,\infty} = C_1 \max_{i\in[d]} \|X_i\|_2$ for all $\phi \in S_{C_1}$. By a union bound, we can write:

$$\begin{aligned}P\left\{\frac{2}{n\hat{\sigma}} \max_{\phi\in S_{C_1}} |\phi^\top X^\top(y - \mu)| \geqslant s\right\} &= P\left\{\frac{2}{n\hat{\sigma}} \max_{\phi\in S_{C_1}^+} |\phi^\top X^\top(y - \mu)| \geqslant s\right\}\\ &\leqslant \sum_{\phi\in S_{C_1}^+} P\left\{\frac{2}{n\hat{\sigma}}|\phi^\top X^\top(y - \mu)| \geqslant s\right\}.\end{aligned}$$

Since $\frac{2}{n\hat{\sigma}}\phi^\top X^\top(y - \mu)$ follows a rescaled t-distribution with r degrees of freedom and there are d terms in the sum on the right-hand side, for $s = s^* := \frac{2t_{r,1-\delta/(2d)}C_1\|X\|_{2,\infty}}{n}$, the probability above is at most δ. Denote $E = \{\omega : \max_{\phi\in S_{C_1}} |\frac{2}{n\hat{\sigma}}\phi^\top X^\top(\omega - \mu)| \leqslant s^*\}$; we have thus shown $P\{y \in E\} \geqslant 1 - \delta$.

We now show that, whenever $y \in E$, stable LASSO with input y is indistinguishable from stable LASSO with input μ. From here on, we fix $y \in E$ and only consider the randomness of the algorithm.

The output of Algorithm 5 can be written as a function of $(\beta_1, \ldots, \beta_{k+1})$, and hence proving that $(\beta_1, \ldots, \beta_{k+1})$ is indistinguishable when computed on y and μ is sufficient to argue that $\hat{\beta}_{\text{LASSO}}$ is indistinguishable on the two inputs, by the post-processing property.

For all $t \leqslant k$, we can write $\beta_{t+1} = g_t(\beta_t, y)$ for some randomized function g_t; in Algorithm 10 we express g_t as an algorithm. If we show $g_t(\beta, y) \approx_{\eta,0} g_t(\beta, \mu)$ for every fixed β such that $\|\beta\|_1 \leqslant C_1$, then we can apply Lemma 4.5.1 to conclude indistinguishability of the whole sequence $(\beta_1, \ldots, \beta_{k+1})$.

Algorithm 10 The g_t subroutine of the stable LASSO algorithm

```
input: $\beta_t, y$
output: $\beta_{t+1}$
  $\forall \phi \in C_1 \cdot \{\pm e_i\}_{i=1}^d$, sample $\xi_{t,\phi} \overset{\text{i.i.d.}}{\sim} \text{Lap}\left(\frac{4t_{r,1-\delta/(2d)} C_1 \|X\|_{2,\infty}}{n\eta}\right)$
  $\forall \phi \in C_1 \cdot \{\pm e_i\}_{i=1}^d$, let $\alpha_\phi = -\frac{2}{n\hat{\sigma}} \phi^\top X^\top (y - X\beta_t) + \xi_{t,\phi}$
  Set $\phi_t = \arg\min_{\phi \in S_{C_1}} \alpha_\phi$
  Set $\beta_{t+1} = (1 - \Delta_t)\beta_t + \Delta_t \phi_t$, where $\Delta_t = \frac{2}{t+1}$
  Return $\beta_{t+1}$
```

Let ϕ_t and ϕ_t^μ denote the minimizers of α_ϕ when the input is y and μ, respectively, and fix an arbitrary point $\phi^* \in S_{C_1}$. Let $\{\xi_{t,\phi}\}_{\phi \in S_{C_1}}$ be independent samples from $\text{Lap}\left(\frac{2s^*}{\eta}\right)$. Denote

$$\xi^* = \arg\max_{\xi} \nabla L(\beta; X, y)^\top \phi^* + \xi \leqslant \nabla L(\beta; X, y)^\top \phi + \xi_{t,\phi}, \forall \phi \in S_{C_1} \setminus \{\phi^*\}.$$

Conditional on $\xi_{t,\phi}$, $\phi \in S_{C_1} \setminus \{\phi^*\}$, we get $\phi_t = \phi^*$ if and only if $\xi_{t,\phi^*} \leqslant \xi^*$.

By the definition of E, we have:

$$\begin{aligned}(\phi^*)^\top \nabla L(\beta; X, \mu) - s^* + \xi^* &\leqslant (\phi^*)^\top \nabla L(\beta; X, y) + \xi^* \\ &\leqslant \phi^\top \nabla L(\beta; X, y) + \xi_{t,\phi} \leqslant \phi^\top \nabla L(\beta; X, \mu) + s^* + \xi_{t,\phi},\end{aligned}$$

for all $\phi \in S_{C_1} \setminus \{\phi^*\}$. As a result, conditional on $\xi_{t,\phi}$, $\phi \in S_{C_1} \setminus \{\phi^*\}$, the event $\xi_{t,\phi^*} \leqslant \xi^* - 2s^*$ implies $\phi_t^\mu = \phi^*$. Thus, we get:

$$\begin{aligned}P\{\phi_t^\mu = \phi^* | \xi_{t,\phi}, \phi \in S_{C_1} \setminus \{\phi^*\}\} &\geqslant P\{\xi_{t,\phi^*} \leqslant \xi^* - 2s^* | \xi_{t,\phi}, \phi \in S_{C_1} \setminus \{\phi^*\}\} \\ &\geqslant e^{-\eta} P\{\xi_{t,\phi^*} \leqslant \xi^* | \xi_{t,\phi}, \phi \in S_{C_1} \setminus \{\phi^*\}\} \\ &= e^{-\eta} P\{\phi_t = \phi^* | \xi_{t,\phi}, \phi \in S_{C_1} \setminus \{\phi^*\}\}.\end{aligned}$$

Applying an expectation to both sides yields

$$P\{\phi_t = \phi^*\} \leqslant e^\eta P\{\phi_t^\mu = \phi^*\},$$

and this is true for all $\phi^* \in S_{C_1}$. Therefore, for all $y \in E$, $\phi_t \approx_{\eta,0} \phi_t^\mu$. By post-processing, this also implies $g_t(\beta, y) \approx_{\eta,0} g_t(\beta, \mu)$, for all β.

By Lemma 4.5.1, we finally conclude that, for all $y \in E$, the output of the stable LASSO algorithm when applied to y is $(\frac{1}{2}k\eta^2 + \sqrt{2k \log(1/\delta)}\eta, \delta)$-indistinguishable from the output implied by the oracle input μ, for all $\delta \in (0, 1)$, or alternatively it is $(k\eta, 0)$-indistinguishable. Since this holds with $1 - \delta$ probability over the choice of y, we see that Algorithm 5 is stable with the desired parameters.

4.5.6 Proof of Proposition 4.3.3 (LASSO utility)

As in the proof of Proposition 4.3.2, we denote the rescaled squared loss by $L(\beta; X, y) := \frac{1}{n\hat{\sigma}}\|y - X\beta\|_2^2$, and by $S_{C_1} := C_1 \cdot \{\pm e_i\}_{i=1}^d$ we denote the set of $2d$ extreme points of the ℓ_1-ball in $\mathbb{R}^d$, scaled by the constraint C_1.

We begin by stating a convergence result for the Frank-Wolfe algorithm due to Jaggi [89], which forms the core of our analysis.

Lemma 4.5.2 ([89]). *Fix $s > 0$ and $\beta_1 \in \mathcal{D} \subseteq \mathbb{R}^d$. Let $(\phi_1, \ldots, \phi_k)$ be a sequence of vectors from $\mathcal{D}$ and let $\beta_{t+1} = (1 - \Delta_t)\beta_t + \Delta_t\phi_t$, for arbitrary $\Delta_t \in [0, 1]$. Define the curvature constant of L as*

$$C_L := \sup_{\beta_1,\beta_2\in\mathcal{D},\gamma\in[0,1],\beta_3=(1-\gamma)\beta_1+\gamma\beta_2} \frac{2}{\gamma^2}(L(\beta_3) - L(\beta_1) - (\beta_3 - \beta_1)^\top \nabla L(\beta_1)).$$

Suppose that for all $t \in [k]$, it holds that: $\phi_t^\top \nabla L(\beta_t) \leqslant \min_{\phi\in C_1\cdot\{\pm e_i\}_{i=1}^d} \phi^\top \nabla L(\beta_t) + \frac{s\Delta_t C_L}{2}$. Then,

$$L(\beta_{k+1}) - \min_{\beta:\|\beta\|_1\leqslant C_1} L(\beta) \leqslant \frac{2C_L}{k+2}(1+s).$$

Denote by $b := \frac{4t_{r,1-\delta/(2d)}C_1\|X\|_{2,\infty}}{n\eta}$ the parameter of the Laplace noise in Algorithm 5. Fix $s > 0$. Denoting by C_L the curvature constant of L, as defined in Lemma 4.5.2, and by applying subexponential concentration of the Laplace distribution, we know:

$$\begin{aligned}
&P\left\{\exists t \in [k] : \phi_t^\top \nabla L(\beta_t; X, y) > \min_{\phi\in S_{C_1}} \phi^\top \nabla L(\beta_t; X, y) + \frac{s\Delta_t C_L}{2}\right\}\\
&\leqslant P\left\{\exists t \in [k] : \max_{\phi\in S_{C_1}} |\xi_{t,\phi}| > \frac{s\Delta_t C_L}{4}\right\}\\
&\leqslant P\left\{\max_{t\in[k],\phi\in S_{C_1}} |\xi_{t,\phi}| > \frac{s\Delta_k C_L}{4}\right\}\\
&\leqslant k|S_{C_1}| \exp\left(-\frac{s\Delta_k C_L}{4b}\right),
\end{aligned}$$

where the last step follows by a union bound. Setting $s = \frac{4b}{\Delta_k C_L}\log(k|S_{C_1}|/\zeta)$ controls this probability to be at most ζ.

We use a standard fact from convex geometry: for any set $S_{\mathcal{D}}$ such that its convex hull is equal to $\mathcal{D}$, it holds that $\min_{\phi\in\mathcal{D}} \phi^\top \nabla L(\beta_t; X, y) = \min_{\phi\in S_{\mathcal{D}}} \phi^\top \nabla L(\beta_t; X, y)$. In our setting, $\mathcal{D} = \{\beta : \|\beta\|_1 \leqslant C_1\}$, and it can be obtained as the convex hull of S_{C_1}.

With this, we can apply Lemma 4.5.2, as well as the fact that $|S_{C_1}| = 2d$, to get that with probability $1 - \zeta$ over the Laplace noise variables:

$$L(\beta_{k+1}; X, y) - \min_{\beta:\|\beta\|_1\leqslant C_1} L(\beta; X, y) \leqslant \frac{2C_L}{k+2} + \frac{8C_L b \log(2kd/\zeta)}{(k+2)\Delta_k C_L}.$$

By the curvature characterization for quadratics due to Clarkson [39], we can bound the curvature constant as

$$C_L \leqslant \frac{1}{n\hat{\sigma}} \max_{\beta,\beta':\|\beta\|_1\leqslant C_1,\|\beta'\|_1\leqslant C_1} \|X(\beta-\beta')\|_2^2 \leqslant \frac{1}{n\hat{\sigma}} \max_{\varphi:\|\varphi\|_1\leqslant 2C_1} \|X\varphi\|_2^2 \leqslant \frac{4}{\hat{\sigma}}\|X\|_\infty^2 C_1^2.$$

Therefore, we can conclude

$$L(\beta_{k+1};X,y) - \min_{\beta:\|\beta\|_1\leqslant C_1} L(\beta;X,y) \leqslant \frac{8\|X\|_\infty^2 C_1^2}{\hat{\sigma}(k+2)} + 4b\log(2kd/\zeta).$$

Further, notice that for all β,β' such that $\max\{\|\beta\|_1,\|\beta'\|_1\}\leqslant C_1$, by Hölder's inequality we have:

$$\begin{aligned}|L(\beta;X,y)-L(\beta';X,y)| &= \left|\frac{1}{n\hat{\sigma}}\|y-X\beta\|_2^2 - \frac{1}{n\hat{\sigma}}\|y-X\beta'\|_2^2\right| \\ &\leqslant \frac{2}{\hat{\sigma}}\|X\|_\infty(\|X\|_\infty C_1 + \|y\|_\infty)\|\beta'-\beta\|_1 := L_1\|\beta'-\beta\|_1 \leqslant 2L_1C_1,\end{aligned}$$

where by L_1 we denote the ℓ_1-Lipschitz constant of the squared loss restricted to the LASSO domain. Now we pick $\zeta = \frac{\gamma}{2C_1L_1}$ for some constant $\gamma > 0$, which gives:

$$\begin{aligned}\mathbb{E}[L(\beta_{k+1};X,y)|y,\hat{\sigma}] - \min_{\beta:\|\beta\|_1\leqslant C_1} L(\beta;X,y) &\leqslant \gamma + \frac{8\|X\|_\infty^2 C_1^2}{\hat{\sigma}(k+2)} + 4b\log(4kdC_1L_1/\gamma) \\ &= \gamma + \frac{8\|X\|_\infty^2 C_1^2}{\hat{\sigma}(k+2)} + \frac{16t_{r,1-\delta/(2d)}C_1\|X\|_{2,\infty}\log(4kdC_1L_1/\gamma)}{n\eta},\end{aligned}$$

where in the last step we use the noise level from Algorithm 5. Now we set $k = \left\lceil \frac{n\|X\|_\infty^2 C_1\eta}{\hat{\sigma}\|X\|_{2,\infty}} \right\rceil$, and get the following utility upper bound:

$$\begin{aligned}&\mathbb{E}[L(\beta_{k+1};X,y)|y,\hat{\sigma}] - \min_{\beta:\|\beta\|_1\leqslant C_1} L(\beta;X,y) \\ &\leqslant \gamma + \frac{8C_1\|X\|_{2,\infty}}{n\eta} + \frac{16t_{r,1-\delta/(2d)}C_1\|X\|_{2,\infty}\log(4kdC_1L_1/\gamma)}{n\eta}.\end{aligned}$$

Note that the above inequality is true for all $\gamma > 0$. After optimizing over γ, the right-hand side reduces to

$$\frac{8C_1\|X\|_{2,\infty}}{n\eta} + \frac{16t_{r,1-\delta/(2d)}C_1\|X\|_{2,\infty}\left(1+\log(kdL_1n\eta/(4t_{r,1-\delta/(2d)}\|X\|_{2,\infty}))\right)}{n\eta}.$$

Using $k \leqslant \frac{2n\|X\|_\infty^2 C_1\eta}{\hat{\sigma}\|X\|_{2,\infty}}$ and the value of L_1, we finally get

$$\begin{aligned}&\mathbb{E}[L(\beta_{k+1};X,y)|y,\hat{\sigma}] - \min_{\beta:\|\beta\|_1\leqslant C_1} L(\beta;X,y) \leqslant \frac{8C_1\|X\|_{2,\infty}}{n\eta} + \\ &\frac{16t_{r,1-\delta/(2d)}C_1\|X\|_{2,\infty}}{n\eta}\left(1+\log\left(\frac{dC_1n^2\eta^2\|X\|_\infty^3(\|X\|_\infty C_1 + \|y\|_\infty)}{2t_{r,1-\delta/(2d)}\|X\|_{2,\infty}^2\hat{\sigma}^2}\right)\right).\end{aligned}$$

Focusing on the relevant parameters, this bound can be simplified as

$$\frac{1}{n}\mathbb{E}[\|y - X\beta_{k+1}\|_2^2 \mid y] - \min_{\beta:\|\beta\|_1 \leqslant C_1} \frac{1}{n}\|y - X\beta\|_2^2 = \tilde{O}\left(\frac{C_1\|X\|_{2,\infty}\log(d)t_{r,1-\delta/(2d)}\sigma}{n\eta}\right).$$

Note that similar guarantees follow without conditioning on y, by taking iterated expectations, applying Jensen's inequality, and using subgaussianity to bound $\mathbb{E}[\|y\|_\infty]$.

4.5.7 Proof of Proposition 4.3.4 (marginal screening stability)

Let $y \sim \mathcal{N}(\mu, \sigma^2 I)$ and define $c_i^\omega := \frac{1}{n\hat{\sigma}} X_i^\top \omega$ for all $\omega \in \mathbb{R}^n$. Let $E = \{\omega : \|c^\omega - c^\mu\|_\infty \leqslant \frac{t_{r,1-\delta/(2d)}\|X\|_{2,\infty}}{n}\}$. First we prove that $P\{y \in E\} \geqslant 1 - \delta$:

$$\begin{aligned} P\left\{\|c^y - c^\mu\|_\infty \geqslant \frac{t_{r,1-\delta/(2d)}\|X\|_{2,\infty}}{n}\right\} &= P\left\{\exists i : \frac{1}{n\hat{\sigma}}|X_i^\top y - X_i^\top \mu| \geqslant \frac{t_{r,1-\delta/(2d)}\|X\|_{2,\infty}}{n}\right\} \\ &= P\left\{\exists i : \left|\frac{X_i^\top(y-\mu)}{\hat{\sigma}}\right| \geqslant t_{r,1-\delta/(2d)}\|X\|_{2,\infty}\right\} \\ &\leqslant d \cdot \frac{\delta}{d} = \delta. \end{aligned}$$

Now we appeal to a similar composition argument as in Proposition 4.3.2. From here on, fix $y \in E$. We will show that the output of stable marginal screening, when applied to y, is indistinguishable from the output of stable marginal screening given the oracle input μ.

The selected model $\hat{M}$ can be written as the output of a composition of k functions $g_t(i_1, \ldots, i_{t-1}, y)$, $t \in [k]$. In particular, the feature "peeled off" at time t, i_t, is equal to $g_t(i_1, \ldots, i_{t-1}, y)$. We show that $g_t(i_1, \ldots, i_{t-1}, y) \approx_{\eta,0} g_t(i_1, \ldots, i_{t-1}, \mu)$ holds true for all fixed $i_1, \ldots, i_{t-1}$. By Lemma 4.5.1, that will imply that the overall selected model under input y and under input μ is indistinguishable as well.

Fix a round $t \in [k]$, as well as an index $i \in \text{res}_t$. Suppose that we add independent draws $\xi_{t,j} \sim \text{Lap}\left(\frac{2t_{r,1-\delta/(2d)}\|X\|_{2,\infty}}{n\eta}\right)$ to each value c_j, where $j \in \text{res}_t$. Define

$$\xi_+^* = \operatorname*{arg\,min}_{\xi \geqslant -c_i^y} c_i^y + \xi > |c_j^y + \xi_{t,j}|, \quad \xi_-^* = \operatorname*{arg\,max}_{\xi < -c_i^y} -c_i^y - \xi > |c_j^y + \xi_{t,j}|, \ \forall j \neq i.$$

Then, $g_t(i_1, \ldots, i_{t-1}, y) = i$ if and only if $\xi_{t,i} \geqslant \xi_+^*$ or $\xi_{t,i} \leqslant \xi_-^*$. Moreover, since $y \in E$, we have

$$\frac{t_{r,1-\delta/(2d)}\|X\|_{2,\infty}}{n} + c_i^\mu + \xi_+^* \geqslant c_i^y + \xi_+^* > |c_j^y + \xi_{t,j}| \geqslant |c_j^\mu + \xi_{t,j}| - \frac{t_{r,1-\delta/(2d)}\|X\|_{2,\infty}}{n},$$

$$\frac{t_{r,1-\delta/(2d)}\|X\|_{2,\infty}}{n} - c_i^\mu - \xi_-^* \geqslant -c_i^y - \xi_-^* > |c_j^y + \xi_{t,j}| \geqslant |c_j^\mu + \xi_{t,j}| - \frac{t_{r,1-\delta/(2d)}\|X\|_{2,\infty}}{n}.$$

Rearranging the terms, we get

$$\frac{2t_{r,1-\delta/(2d)}\|X\|_{2,\infty}}{n} + c_i^\mu + \xi_+^* \geqslant |c_j^\mu + \xi_{t,j}|, \ \frac{2t_{r,1-\delta/(2d)}\|X\|_{2,\infty}}{n} - c_i^\mu - \xi_-^* \geqslant |c_j^\mu + \xi_{t,j}|.$$

Thus, if $\xi_{t,i} \geqslant \xi_+^* + \frac{2t_{r,1-\delta/(2d)}\|X\|_{2,\infty}}{n}$ or $\xi_{t,i} \leqslant \xi_-^* - \frac{2t_{r,1-\delta/(2d)}\|X\|_{2,\infty}}{n}$, then $i = g_t(i_1, \dots, i_{t-1}, \mu)$ if the noise levels are $(\xi_{t,1}, \dots, \xi_{t,i}, \dots, \xi_{t,d})$. Finally, for fixed $y \in E$, we have

$$\begin{aligned}
&P\left\{g_t(i_1, \dots, i_{t-1}, \mu) = i | \{\xi_{t,j}\}_{j \neq i}\right\} \\
&\geqslant P\left\{\xi_{t,i} \geqslant \xi_+^* + \frac{2t_{r,1-\delta/(2d)}\|X\|_{2,\infty}}{n} \middle| \{\xi_{t,j}\}_{j\neq i}\right\} \\
&+ P\left\{\xi_{t,i} \leqslant \xi_-^* - \frac{2t_{r,1-\delta/(2d)}\|X\|_{2,\infty}}{n} \middle| \{\xi_{t,j}\}_{j\neq i}\right\} \\
&\geqslant e^{-\eta} P\left\{\xi_{t,i} \geqslant \xi_+^* | \{\xi_{t,j}\}_{j\neq i}\right\} + e^{-\eta} P\left\{\xi_{t,i} \leqslant \xi_-^* | \{\xi_{t,j}\}_{j\neq i}\right\} \\
&= e^{-\eta} P\left\{g_t(i_1, \dots, i_{t-1}, y) = i | \{\xi_{t,j}\}_{j\neq i}\right\}.
\end{aligned}$$

Multiplying by e^η and applying the law of iterated expectations completes the proof that $g_t(i_1, \dots, i_{t-1}, y) \approx_{\eta,0} g_t(i_1, \dots, i_{t-1}, \mu)$ for all $y \in E$.

Finally, by Lemma 4.5.1 we conclude that for all fixed $y \in E$, the output of stable marginal screening under input y and under the oracle input μ is $(\frac{1}{2}k\eta^2 + \sqrt{2k\log(1/\delta)}\eta, \delta)$-indistinguishable for all $\delta \in (0,1)$, or alternatively $(k\eta, 0)$-indistinguishable. Since this holds with $1-\delta$ probability over the choice of y, we see that stable marginal screening satisfies stability with the desired parameters.

4.5.8 Proof of Proposition 4.3.5 (marginal screening utility)

Fix $s > 0$. Taking a union bound, we get:

$$P\left\{\max_{j\in[k]} |c_{m_j}| - |c_{i_j}| \geqslant s \middle| y\right\} \leqslant \sum_{j=1}^k P\left\{|c_{m_j}| - |c_{i_j}| \geqslant s \middle| y\right\}.$$

At the time when i_j is chosen, exactly $j-1$ items have been selected; therefore, at least one of $m_1, \dots, m_j$ has still not been selected. The event that $|c_{m_j}| - |c_{i_j}| \geqslant s$ implies that i_j "beat" one of $m_1, \dots, m_j$, which further implies that $\max_{i\in[d]} |\xi_{j,i}| \geqslant \frac{s}{2}$. By a union bound, this happens with probability at most $d\exp(-sn\eta/(4t_{r,1-\delta/(2d)}\|X\|_{2,\infty}))$. Putting everything together, we get

$$\sum_{j=1}^k P\left\{|c_{m_j}| - |c_{i_j}| \geqslant s \middle| y\right\} \leqslant kd\exp\left(-\frac{sn\eta}{4t_{r,1-\delta/(2d)}\|X\|_{2,\infty}}\right).$$

Plugging in $s = \frac{4t_{r,1-\delta/(2d)}\log(dk/\delta')\|X\|_{2,\infty}}{n\eta}$ completes the proof.

4.5.9 Proof of Theorem 4.4.3

The proof essentially follows by applying Theorem 4.4.2. The bulk of the proof applies to both the problem of inference on the winner and the file-drawer problem. Towards the end we specialize the analysis to the individual problems.

We let

$$A_\nu(\mu) = \{y : \|y-\mu\|_\infty \leqslant q^\nu([m])\}\,.$$

The validity of this set follows directly by the definition of $q^\nu([m])$:

$$P_\mu\left\{\max_{i\in[m]}|y_i-\mu_i| \leqslant q^\nu([m])\right\} = P_0\left\{\max_{i\in[m]}|Z_i| \leqslant q^\nu([m])\right\} \geqslant 1-\alpha.$$

Using the above choice of $A_\nu(\mu)$ we can write

$$\Gamma_\nu(\mu) = \cup_{y:\|y-\mu\|_\infty\leqslant q^\nu([m])}\widehat{\Gamma}(y),$$

and thus

$$\begin{aligned}\widehat{\Gamma}_\nu^+ &= \cup_{\mu:\|y-\mu\|_\infty\leqslant q^\nu([m])}\Gamma_\nu(\mu)\\ &= \cup_{\mu:\|y-\mu\|_\infty\leqslant q^\nu([m])}\cup_{y':\|y'-\mu\|_\infty\leqslant q^\nu([m])}\widehat{\Gamma}(y')\\ &= \cup_{y':\|y'-y\|_\infty\leqslant 2q^\nu([m])}\widehat{\Gamma}(y').\end{aligned}$$

In words, the set $\widehat{\Gamma}_\nu^+$ is the set of all selections obtained by perturbing the entries in y by at most $2q^\nu([m])$ in ℓ_∞-norm.

Now we specialize the analysis to the two selection problems.

In the problem of inference on the winner, the selected inferential target is indexed by $\widehat{\Gamma}(y) = \{\hat{\gamma}(y)\}$. The most favorable perturbation y' for an index $j\in[m]$ to be selected is obtained by taking $y'_j = y_j + 2q^\nu([m])$ and $y'_k = y_k - 2q^\nu([m])$ for $k\neq j$; therefore, $\widehat{\Gamma}_\nu^+ = \{\gamma\in[m] : y_\gamma \geqslant y_{\hat{\gamma}} - 4q^\nu([m])\}$ is the set of plausible selections.

In the file-drawer problem, the selected targets are $\widehat{\Gamma}(y) = \{\gamma\in[m] : y_\gamma \geqslant T\}$. The indices that could fall in this set given a $2q^\nu([m])$ perturbation around y are $\widehat{\Gamma}_\nu^+ = \{\gamma\in[m] : y_\gamma \geqslant T - 2q^\nu([m])\}$.

Therefore, in both cases we have identified $\widehat{\Gamma}_\nu^+$. The final statement follows by applying Theorem 4.4.2.

4.5.10 Proof of Theorem 4.4.4

By following virtually the same argument as in Theorem 4.4.3, we can conclude that $\widehat{\Gamma}_\nu^+$ is the set of all plausible selections, for both the problem of inference on the winner and the file-drawer problem. The final statement follows by applying Theorem 4.4.1, together with a Bonferroni correction over $\widehat{\Gamma}_\nu^+$.

4.5.11 Proof of Corollary 4.4.1

We argue that $\widehat{\mathcal{V}}_\nu^+$ is the set of plausible targets $\widehat{\Gamma}_\nu^+$ when the acceptance region is chosen as

$$A_\nu(\mu) = \left\{y : \|X^\top y - X^\top \mu\|_\infty \leqslant q^\nu(\{X_j\}_{j=1}^d)\right\}.$$

After this is established, the result follows directly from Theorem 4.4.1.

First, the validity of $A_\nu(\mu)$ follows by the definition of q^ν:

$$P_\mu\left\{\|X^\top y - X^\top \mu\|_\infty > q^\nu(\{X_j\}_{j=1}^d)\right\} = P_0\left\{\max_{j\in[d]} |X_j^\top Z| > q^\nu(\{X_j\}_{j=1}^d)\right\} \leqslant \nu,$$

where $Z \sim P_0$. Therefore, $P_\mu\{y \in A_\nu(\mu)\} \geqslant 1-\nu$.

We can write

$$\Gamma_\nu(\mu) = \cup_{y:\|X^\top y - X^\top\mu\|_\infty \leqslant q^\nu(\{X_j\}_{j=1}^d)} \widehat{\Gamma}(y),$$

and thus

$$\begin{aligned}\widehat{\Gamma}_\nu^+ &= \cup_{\mu:\|X^\top y - X^\top\mu\|_\infty \leqslant q^\nu(\{X_j\}_{j=1}^d)} \Gamma_\nu(\mu)\\ &= \cup_{\mu:\|X^\top y - X^\top\mu\|_\infty \leqslant q^\nu(\{X_j\}_{j=1}^d)} \cup_{y':\|X^\top y' - X^\top\mu\|_\infty \leqslant q^\nu(\{X_j\}_{j=1}^d)} \widehat{\Gamma}(y')\\ &= \cup_{y':\|X^\top y - X^\top y'\|_\infty \leqslant 2q^\nu(\{X_j\}_{j=1}^d)} \widehat{\Gamma}(y').\end{aligned}$$

Since $\widehat{\Gamma}(y) = \left\{(j, \hat{M}(y)) : j \in \hat{M}(y)\right\}$, we finally have

$$\begin{aligned}\widehat{\Gamma}_\nu^+ &= \cup_{y':\|X^\top y - X^\top y'\|_\infty \leqslant 2q^\nu(\{X_j\}_{j=1}^d)} \left\{(j, \hat{M}(y')) : j \in \hat{M}(y')\right\}\\ &= \left\{(j, M) : j \in M, M \in \widehat{\mathcal{M}}_\nu^+\right\}.\end{aligned}$$

Therefore, by Theorem 4.4.1 it suffices to take a simultaneous correction over $\widehat{\Gamma}_\nu^+$. The set $\widehat{\mathcal{V}}^+$ is the set of contrasts that ensures simultaneously valid inference for $\{\theta_{j\cdot M}\}_{M\in\widehat{\mathcal{M}}_\nu^+}$ (see, e.g., Theorem 4.1 in [12]).

4.5.12 Proof of Lemma 4.4.2

First, we argue that the condition $|X_j^\top(y - X_M\beta_{(M,s)}(y))| < \lambda - s_\nu(1 + \|X_j^\top X_M(X_M^\top X_M)^{-1}\|_1)$ is equivalent to

$$\max_{y'\in\mathcal{B}_\nu^\infty} |X_j^\top(y' - X_M\beta_{(M,s)}(y'))| < \lambda.$$

This follows because

$$\begin{aligned}&\max_{y'\in\mathcal{B}_\nu^\infty} |X_j^\top(y' - X_M\beta_{(M,s)}(y'))|\\ &= \max_{y'\in\mathcal{B}_\nu^\infty} \left|X_j^\top(y - X_M\beta_{(M,s)}(y)) + X_j^\top\left(y' - y - X_M\beta_{(M,s)}(y') + X_M\beta_{(M,s)}(y)\right)\right|\\ &= \max_{y'\in\mathcal{B}_\nu^\infty} \left|X_j^\top(y - X_M\beta_{(M,s)}(y)) + X_j^\top\left(I - X_M(X_M^\top X_M)^{-1}X_M^\top\right)(y'-y)\right|.\end{aligned}$$

Now notice that by the definition of $\mathcal{B}_\nu^\infty$ we have

$$\begin{aligned}\max_{y'\in\mathcal{B}_\nu^\infty} X_j^\top(I - X_M(X_M^\top X_M)^{-1}X_M^\top)(y'-y) &= \max_{|z_j|\leqslant s_\nu, \|z_M\|_\infty \leqslant s_\nu} z_j - X_j^\top X_M(X_M^\top X_M)^{-1}z_M \\ &= s_\nu + s_\nu\|X_j^\top X_M(X_M^\top X_M)^{-1}\|_1,\end{aligned}$$

which follows by the duality between the ℓ_1- and ℓ_∞-norms. Similarly we have

$$\min_{y'\in\mathcal{B}_\nu^\infty} X_j^\top(I - X_M(X_M^\top X_M)^{-1}X_M^\top)(y'-y) = -s_\nu - s_\nu\|X_j^\top X_M(X_M^\top X_M)^{-1}\|_1.$$

Thus, we see that

$$\begin{aligned}&\max_{y'\in\mathcal{B}_\nu^\infty} \left|X_j^\top(y - X_M\beta_{(M,s)}(y)) + X_j^\top(I - X_M(X_M^\top X_M)^{-1}X_M^\top)(y'-y)\right| \\ &\quad = \left|X_j^\top(y - X_M\beta_{(M,s)}(y))\right| + s_\nu\left(1 + \|X_j^\top X_M(X_M^\top X_M)^{-1}\|_1\right)\end{aligned}$$

Putting everything together, we have shown that

$$\max_{y'\in\mathcal{B}_\nu^\infty} |X_j^\top(y' - X_M\beta_{(M,s)}(y'))| = \left|X_j^\top(y - X_M\beta_{(M,s)}(y))\right| + s_\nu\left(1 + \|X_j^\top X_M(X_M^\top X_M)^{-1}\|_1\right).$$

Therefore, the screening rule in Lemma 4.4.2 is equivalent to

$$\max_{y'\in\mathcal{B}_\nu^\infty} \left|X_j^\top(y' - X_M\beta_{(M,s)}(y'))\right| < \lambda. \tag{4.11}$$

We argue that the condition in Eq. (4.11) implies that there cannot exist a pair $(M', s') \in \mathcal{B}(M, s)$ such that $M' = M \cup \{j\}$. Indeed, if this were true, then there must exist a point $y' \in \mathcal{B}_\nu^\infty$ on the boundary between the two corresponding polyhedra. Given the polyhedral characterization of Lee et al. [106], this point must satisfy

$$X_j^\top(I - X_M(X_M^\top X_M)^{-1}X_M^\top)y' = \lambda(1 - X_j^\top(X_M^\top)^+ s).$$

By rearranging, we see that this equality is equivalent to

$$X_j^\top\left(y' - X_M(X_M^\top X_M)^{-1}(X_M^\top y' - \lambda s)\right) = \lambda. \tag{4.12}$$

The left-hand side is equal to $X_j^\top(y' - X_M\beta_{(M,s)}(y'))$; therefore, condition (4.12) contradicts condition (4.11), and thus we can conclude that $M \cup \{j\}$ cannot be the model of any neighboring model-sign pair $(M', s') \in \mathcal{B}(M, s)$.

4.5.13 Proof of Lemma 4.4.3

The proof proceeds similarly to the proof of Lemma 4.4.2. Fix $j \in M$. First, we argue that the condition $|\beta_{j\cdot(M,s)}(y)| > s_\nu\|e_{j\cdot(M,s)}^\top(X_M^\top X_M)^{-1}\|_1$ is equivalent to

$$\min_{y'\in\mathcal{B}_\nu^\infty} |\beta_{j\cdot(M,s)}(y')| > 0. \tag{4.13}$$

This follows by writing

$$\begin{aligned}\min_{y' \in \mathcal{B}_\nu^\infty} |\beta_{j\cdot(M,s)}(y')| &= \min_{y' \in \mathcal{B}_\nu^\infty} |\beta_{j\cdot(M,s)}(y) + \beta_{j\cdot(M,s)}(y') - \beta_{j\cdot(M,s)}(y)| \\ &= \min_{y' \in \mathcal{B}_\nu^\infty} |\beta_{j\cdot(M,s)}(y) + e_{j\cdot(M,s)}^\top (X_M^\top X_M)^{-1} X_M^\top (y' - y)|.\end{aligned}$$

By the definition of $\mathcal{B}_\nu^\infty$, we can write

$$\max_{y' \in \mathcal{B}_\nu^\infty} e_{j\cdot(M,s)}^\top (X_M^\top X_M)^{-1} X_M^\top (y' - y) = \max_{\|z_M\|_\infty \leqslant s_\nu} e_{j\cdot(M,s)}^\top (X_M^\top X_M)^{-1} z_M = s_\nu \|e_{j\cdot(M,s)}^\top (X_M^\top X_M)^{-1}\|_1,$$

and similarly $\min_{y' \in \mathcal{B}_\nu^\infty} e_{j\cdot(M,s)}^\top (X_M^\top X_M)^{-1} X_M^\top (y' - y) = -s_\nu \|e_{j\cdot(M,s)}^\top (X_M^\top X_M)^{-1}\|_1$. Putting everything together, we see that $\min_{y' \in \mathcal{B}_\nu^\infty} |\beta_{j\cdot(M,s)}(y')| > 0$ implies $s_\nu \|e_{j\cdot(M,s)}^\top (X_M^\top X_M)^{-1}\|_1 < |\beta_{j\cdot(M,s)}(y)|$, and vice versa.

Now we argue that condition (4.13) implies that variable j cannot exit the model in any of the neighboring polyhedra within $\mathcal{B}_\nu^\infty$. If it can, then the Lee et al. [106] characterization implies that there exists a point $y' \in \mathcal{B}_\nu^\infty$ on the boundary between the respective polyhedra such that $e_{j\cdot(M,s)}^\top (X_M^\top X_M)^{-1} y' = \lambda e_{j\cdot(M,s)}^\top (X_M^\top X_M)^{-1} s$. By rearranging, we can rewrite this equality as $e_{j\cdot(M,s)}^\top \beta_{j\cdot(M,s)}(y') = 0$, which contradicts condition (4.13).

Chapter 5

Prediction-Powered Inference

Machine-learning algorithms are increasingly employed as black-box systems that supply predictions to augment or supplant costly experimental measurements. Such machine-learning systems, generally trained on experimental data, can be used to generate predictions for large numbers of entities that were not studied experimentally. For example, predictions of three-dimensional structure can be made for the entire catalog of known proteins via AlphaFold [92]. Such predictions hold out the promise of increasing the pace and scope of scientific inquiry, particularly in domains where large numbers of entities need to be screened, such as in assessment of molecular activity, tumor prognoses, or micro-climatic modeling. Moreover, there is a cumulative effect—chains of predictions can feed further predictions. As prediction-based scientific inquiry becomes increasingly common, an urgent agenda item is to assess its support in terms of basic principles of statistical inference.

In this chapter we ask whether it is possible to get the best of both worlds—to exploit predictions from a machine-learning system while still providing guarantees of statistical validity. We study this question in a general setting with n data points accompanied by gold-standard labels and N unlabeled data points whose labels are predicted by a machine-learning model. We assume that N is much larger than n. We take the prediction model as pre-existing, as in cases like AlphaFold, where a model was trained offline perhaps at great expense and with massive amounts of data. We consider a scientist who wishes to use predictions from the model to perform inference. The scientist's goal is not to replace the experimental data with predictions, but rather to leverage the immense number of predictions to improve their confidence in a scientific conclusion.

We present *prediction-powered inference*, a framework that provides an affirmative answer to the question of whether predictions can improve inferential quality without sacrificing rigorous validity guarantees. Rather than using predictions as raw data, prediction-powered inference uses the small gold-standard data set of paired features and labels to estimate a mathematical object that we refer to as the *rectifier*. The rectifier makes it possible to transform parameter estimates based on predictions into a statistically valid confidence set.

The material in this chapter is based on a work co-authored with Anastasios Angelopoulos, Stephen Bates, Clara Fannjiang, and Michael I. Jordan [3].

5.1 General principle

We now overview prediction-powered inference. The goal is to estimate a quantity θ^*, such as the mean or median value of a random outcome. Towards this goal, we have access to a small gold-standard data set of paired features and outcomes, $(X, Y) = \big((X_1, Y_1), \ldots, (X_n, Y_n)\big)$, as well as the features from a large unlabeled data set, $(\widetilde{X}, \widetilde{Y}) = \big((\widetilde{X}_1, \widetilde{Y}_1), \ldots, (\widetilde{X}_N, \widetilde{Y}_N)\big)$, where we do not observe the true outcomes $\widetilde{Y}_1, \ldots, \widetilde{Y}_N$. We care about the case where $N \gg n$. For both data sets, we have predictions of the outcome made by a machine-learning algorithm f, denoted $f(X) = (f(X_1), \ldots, f(X_n))$ and $f(\widetilde{X}) = (f(\widetilde{X}_1), \ldots, f(\widetilde{X}_N))$.

Prediction-powered inference builds confidence intervals that are guaranteed to contain θ^*. Imagine we have an estimator $\hat{\theta}$ of θ^*. One feasible but naive way to estimate θ^*, which we call the *imputation approach*, is to treat the predictions as gold-standard outcomes and compute $\tilde{\theta}^f = \hat{\theta}(\widetilde{X}, f(\widetilde{X}))$. If the predictions are accurate, meaning $f(\widetilde{X}_i) \approx \widetilde{Y}_i$, then $\tilde{\theta}^f$ is close to θ^*. However, $\tilde{\theta}^f$ will generally be biased due to errors in the predictions. Instead, our key idea is to use the gold-standard data set to quantify how the prediction errors affect the imputed estimate, and then construct a confidence set for θ^* by adjusting for this effect.

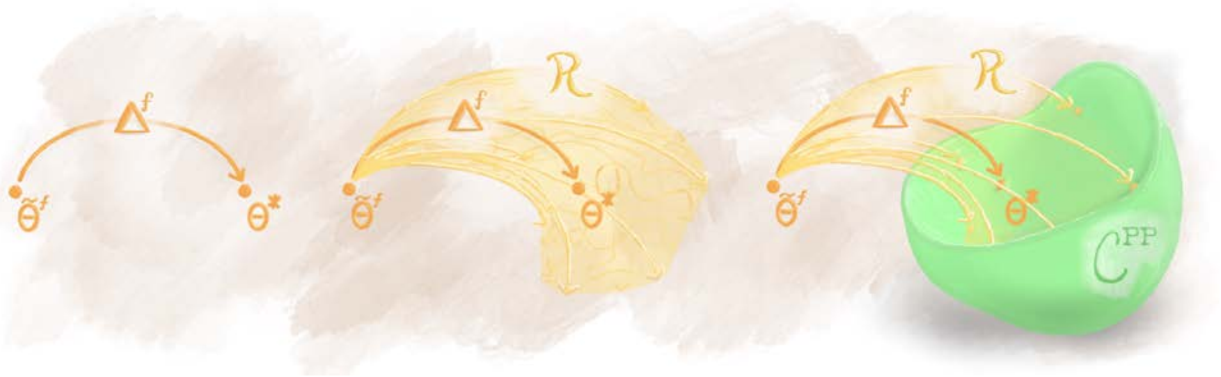
More systematically, the first step is to introduce a problem-specific measure of prediction error called the *rectifier*, denoted as $\boldsymbol{\Delta}^f$. The rectifier captures how errors in the predictions lead to bias in $\tilde{\theta}^f$. Intuitively, $\boldsymbol{\Delta}^f$ recovers θ^* by "rectifying" $\tilde{\theta}^f$. The appropriate rectifier depends on the estimand of interest θ^*, and we show how to derive it for a broad class of estimands. Next, we use the gold-standard data to construct a confidence set for the rectifier, $\mathcal{R}$. Finally, we form a confidence set for θ^* by taking $\tilde{\theta}^f$ and rectifying it with each possible value in the set $\mathcal{R}$. The collection of these rectified values is the prediction-powered confidence set, $\mathcal{C}^{\mathrm{PP}}$, which is guaranteed to contain θ^* with high probability.

Prediction-powered inference leads to powerful and provably valid confidence intervals and p-values for a broad class of statistical problems, enabling researchers to reliably incorporate machine learning into their analyses. We provide practical algorithms for constructing prediction-powered confidence intervals for means, quantiles, modes, linear and logistic regression coefficients, as well as other inferential targets. For conciseness, our technical statements and algorithms will focus on constructing confidence intervals; however, note that through the duality between confidence intervals and hypothesis tests, our intervals directly imply valid prediction-powered p-values and hypothesis tests as well.

5.1.1 Further preliminaries

We use $(X, Y) \in (\mathcal{X} \times \mathcal{Y})^n$ to denote the labeled data set, where $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$. We use the terms "labeled" and "gold-standard" interchangeably. We use analogous notation for the unlabeled data set, $(\widetilde{X}, \widetilde{Y}) \in (\mathcal{X} \times \mathcal{Y})^N$, where the outcomes $\widetilde{Y}$ are not observed. For now we assume that (X, Y) and $(\widetilde{X}, \widetilde{Y})$ are independently and identically distributed samples from a common distribution, $\mathbb{P}$. We generalize our results to settings with distribution shift later on. By θ^* we denote the estimand of interest, which

Prediction-Powered Inference



1. Define rectifier

Define the *rectifier*, $\mathbf{\Delta}^f$, a measure of prediction error.

2. Rectifier confidence set

With labeled data, create $\mathcal{R}$, a confidence set for the rectifier.

3. Prediction-powered confidence set

Construct confidence set, $\mathcal{C}^{\mathrm{PP}}$, by rectifying $\tilde{\theta}^f$ with each value in the set $\mathcal{R}$.

will typically be an underlying property of $\mathbb{P}$, such as the mean outcome.

Next, we have a prediction rule, $f : \mathcal{X} \to \mathcal{Y}$, that is independent of the observed data. For example, it may have been trained on other data independent from both the labeled and the unlabeled data. We let $f_i = f(X_i)$ denote the predictions for the labeled data and $\tilde{f}_i = f(\widetilde{X}_i)$ denote the predictions for the unlabeled data. Slightly abusing notation, we let $f = (f_1, \ldots, f_n)$ and $\tilde{f} = (\tilde{f}_1, \ldots, \tilde{f}_N)$. We will treat $X, Y, \widetilde{X}, \widetilde{Y}, f, \tilde{f}$ as vectors and matrices where appropriate.

Our key conceptual innovation is the *rectifier* $\mathbf{\Delta}^f$—a measure of the prediction rule's accuracy. We formally define the rectifier in Section 5.2. We use $\hat{\mathbf{\Delta}}^f$ to denote an estimate of the rectifier based on labeled data, which we call the empirical rectifier.

5.1.2 Warmup: Mean estimation

Before presenting our main results, we use the example of mean estimation to build intuition. Our goal is to give a valid confidence interval for the average outcome, $\theta^* = \mathbb{E}[Y_1]$. The classical estimate of θ^* is the sample average of the outcomes on the labeled data set, $\hat{\theta}^{\mathrm{class}} = \frac{1}{n}\sum_{i=1}^n Y_i$. We construct a prediction-powered estimate, $\hat{\theta}^{\mathrm{PP}}$, and show that it leads to tighter confidence intervals than $\hat{\theta}^{\text{class}}$ if the prediction rule is accurate. Consider

$$\hat{\theta}^{\text{PP}} = \underbrace{\frac{1}{N}\sum_{i=1}^{N} \tilde{f}_i}_{\tilde{\theta}^f} - \underbrace{\frac{1}{n}\sum_{i=1}^{n}(f_i - Y_i)}_{\hat{\boldsymbol{\Delta}}^f}. \tag{5.1}$$

The key idea is that if the predictions are accurate, we have $\hat{\boldsymbol{\Delta}}^f \approx 0$ and $\hat{\theta}^{\text{PP}} \approx \frac{1}{N}\sum_{i=1}^{N} \widetilde{Y}_i$, which has a much lower variance than $\hat{\theta}^{\text{class}}$ since $N \gg n$.

Notice $\hat{\theta}^{\text{PP}}$ is unbiased for θ^* and it is a sum of two independent terms. Thus, we can construct 95% confidence intervals for θ^* as

$$\underbrace{\hat{\theta}^{\text{PP}} \pm 1.96\sqrt{\frac{\hat{\sigma}^2_{f-Y}}{n} + \frac{\hat{\sigma}^2_{\tilde{f}}}{N}}}_{\text{prediction-powered interval}} \quad \text{or} \quad \underbrace{\hat{\theta}^{\text{class}} \pm 1.96\sqrt{\frac{\hat{\sigma}^2_{Y}}{n}}}_{\text{classical interval}}, \tag{5.2}$$

where $\hat{\sigma}^2_Y$, $\hat{\sigma}^2_{f-Y}$, and $\hat{\sigma}^2_{\tilde{f}}$ are the estimated variances of the Y_i, $f_i - Y_i$, and $\tilde{f}_i$, respectively. The prediction-powered confidence interval is better than the classical interval when the model is good. Because $N \gg n$, the width of the prediction-powered interval is primarily determined by the term $\hat{\sigma}^2_{f-Y}$. Furthermore, when the model has small errors, we have $\hat{\sigma}^2_{f-Y} \ll \hat{\sigma}^2_Y$. Thus, the width of the prediction-powered interval will be smaller than the width of the classical interval. This estimator is known in the literature as the difference estimator, closely related to generalized regression estimators [29]. This variance reduction is why prediction-powered confidence intervals are smaller than their classical counterparts in a broad range of settings beyond mean estimation.

5.2 Main theory: Convex estimation

Our main contribution is a technique for inference on estimands that can be expressed as the solution to a *convex optimization problem.* In addition to means, this includes medians, other quantiles, linear and logistic regression coefficients, and many other quantities. Formally, we consider estimands of the form

$$\theta^* = \underset{\theta \in \mathbb{R}^p}{\arg\min}\ \mathbb{E}\left[\ell_\theta(X_1, Y_1)\right], \tag{5.3}$$

for a loss function $\ell_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that is convex in $\theta \in \mathbb{R}^p$, for some $p \in \mathbb{N}$. Throughout, we take the existence of θ^* as given. If the minimizer is not unique, our method will return a confidence set guaranteed to contain all minimizers. Under mild conditions, convexity ensures that θ^* can also be expressed as the value solving

$$\mathbb{E}\left[g_{\theta^*}(X_1, Y_1)\right] = 0, \tag{5.4}$$

where $g_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^p$ is a subgradient of ℓ_θ with respect to θ. We will call convex estimation problems where θ^* satisfies (5.4) nondegenerate, and we will later discuss mild conditions that ensure this regularity.

Defining the rectifier. Following the outline in Section 5.1, the first step in prediction-powered inference is to define a rectifier. As in the mean estimation case, the rectifier captures a notion of prediction error. In the general setting of convex estimation problems, the relevant notion of error is the bias of the subgradient g_θ computed using the predictions:

$$\boldsymbol{\Delta}^f(\theta) = \mathbb{E}\left[g_\theta(X_1, Y_1) - g_\theta(X_1, f_1)\right]. \tag{5.5}$$

Rectifier confidence set. The second step is to create a confidence set for the rectifier, $\mathcal{R}_\delta(\theta)$, satisfying

$$P\left(\boldsymbol{\Delta}^f(\theta) \in \mathcal{R}_\delta(\theta)\right) \geqslant 1 - \delta. \tag{5.6}$$

Because the rectifier is an expectation for each θ, $\mathcal{R}_\delta(\theta)$ can be constructed using standard, off-the-shelf confidence intervals for the mean.

Prediction-powered confidence set. The final step is to form a confidence set for θ^*. We do so by combining $\mathcal{R}_\delta(\theta)$ with a term that accounts for finite-sample fluctuations due to having N samples. In particular, for every θ, we want a confidence set $\mathcal{T}_{\alpha-\delta}(\theta)$ for $\mathbb{E}[g_\theta(X_1, f_1)]$, satisfying

$$P\left(\mathbb{E}[g_\theta(X_1, f_1)] \in \mathcal{T}_{\alpha-\delta}(\theta)\right) \geqslant 1 - (\alpha - \delta).$$

Again, since $\mathbb{E}[g_\theta(X_1, f_1)]$ is a mean, constructing $\mathcal{T}_{\alpha-\delta}(\theta)$ is easy and can be done with off-the-shelf tools.

We put all the steps together in Theorem 5.2.1.

Theorem 5.2.1 (Convex estimation)**.** *Suppose that the convex estimation problem is nondegenerate as in* (5.4)*. Fix $\alpha \in (0,1)$ and $\delta \in (0, \alpha)$. Suppose that, for any $\theta \in \mathbb{R}^p$, we can construct $\mathcal{R}_\delta(\theta)$ and $\mathcal{T}_{\alpha-\delta}(\theta)$ satisfying*

$$P\left(\boldsymbol{\Delta}^f(\theta) \in \mathcal{R}_\delta(\theta)\right) \geqslant 1 - \delta; \quad P\left(\mathbb{E}[g_\theta(X_1, f_1)] \in \mathcal{T}_{\alpha-\delta}(\theta)\right) \geqslant 1 - (\alpha - \delta).$$

Let $\mathcal{C}_\alpha^{\mathrm{PP}} = \{\theta : 0 \in \mathcal{R}_\delta(\theta) + \mathcal{T}_{\alpha-\delta}(\theta)\}$, where $+$ denotes the Minkowski sum.[1] *Then,*

$$P(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}) \geqslant 1 - \alpha. \tag{5.7}$$

This result means that we can construct a valid confidence set for θ^*, without assumptions about the data distribution or the machine-learning model, for any nondegenerate convex estimation problem. We also present an asymptotic counterpart of Theorem 5.2.1 in Section 5.6.2.

Most practical problems are nondegenerate (5.4). For example, if the loss is differentiable for all $\theta \in \mathbb{R}^p$, then the problem is immediately nondegenerate. Furthermore, if the data distribution does not have point masses and, for every θ, $\ell_\theta(x, y)$ is nondifferentiable only for a measure-zero set of (x, y) pairs, then the problem is again nondegenerate.

[1] The Minkowski sum of two sets A and B is equal to $\{a + b : a \in A, b \in B\}$.

We have focused on convex estimation problems, since this is a broad class of estimands addressed by prediction-powered inference. Nonetheless, we highlight that the general principles for prediction-powered inference from Section 5.1 are applicable more broadly, and lead to additional results and algorithms for other estimands and some forms of distribution shift; see Section 5.4 for such extensions.

5.2.1 Algorithms

In this section we present prediction-powered algorithms for several canonical inference problems. The algorithms rely on confidence intervals derived from the central limit theorem. We implicitly assume the standard, mild regularity conditions required for the asymptotic validity of such intervals. In the algorithms we use $z_{1-\delta}$ to denote the $1-\delta$ quantile of the standard normal distribution, for $\delta \in (0,1)$.

Mean estimation. We begin by returning to the problem of mean estimation:

$$\theta^* = \mathbb{E}[Y_1]. \tag{5.8}$$

The mean can alternatively be expressed as the solution to a convex optimization problem by writing it as the minimizer of the average squared loss:

$$\theta^* = \arg\min_{\theta \in \mathbb{R}} \mathbb{E}[\ell_\theta(Y_1)] = \arg\min_{\theta \in \mathbb{R}} \mathbb{E}\left[\frac{1}{2}(Y_1 - \theta)^2\right].$$

The squared loss $\ell_\theta(y)$ is differentiable, with gradient equal to $g_\theta(y) = \theta - y$. Applying this in the definition of the rectifier (5.5), we get $\boldsymbol{\Delta}^f(\theta) \equiv \boldsymbol{\Delta}^f = \mathbb{E}[f_1 - Y_1]$. Note that this rectifier has no dependence on θ. We provide an explicit algorithm for prediction-powered mean estimation and its guarantee in Algorithm 11 and Proposition 5.2.1, respectively.

Proposition 5.2.1 (Mean estimation). *Let θ^* be the mean outcome* (5.8). *Then, the prediction-powered confidence interval in Algorithm 11 has valid coverage:* $\liminf_{n,N\to\infty} P\left(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}\right) \geqslant 1-\alpha$.

Quantile estimation. We now turn to quantile estimation. For a pre-specified level $q \in (0,1)$, we wish to estimate the q-quantile of the outcome distribution:

$$\theta^* = \min\left\{\theta : P\left(Y_1 \leqslant \theta\right) \geqslant q\right\}. \tag{5.9}$$

To simplify the exposition, we assume that the distribution of Y_1 does not have point masses; this ensures that the problem is nondegenerate (5.4), though it is possible to generalize beyond this setting with a standard construction. It is well known [101] that the q-quantile can be expressed in variational form as

$$\theta^* = \arg\min_{\theta \in \mathbb{R}}\ \mathbb{E}\left[\ell_\theta(Y_1)\right] = \arg\min_{\theta \in \mathbb{R}}\ \mathbb{E}\left[q(Y_1 - \theta)\mathbf{1}\left\{Y_1 > \theta\right\} + (1-q)(\theta - Y_1)\mathbf{1}\left\{Y_1 \leqslant \theta\right\}\right], \tag{5.10}$$

where ℓ_θ is called the quantile loss (or "pinball" loss). The quantile loss has subgradient $g_\theta(y) = -q\mathbf{1}\{y > \theta\} + (1-q)\mathbf{1}\{y \leqslant \theta\} = -q + \mathbf{1}\{y \leqslant \theta\}$. Plugging the expression for $g_\theta(y)$ into the definition (5.5), we get the relevant rectifier: $\boldsymbol{\Delta}^f(\theta) = P(Y_1 \leqslant \theta) - P(f_1 \leqslant \theta) = \mathbb{E}\left[\mathbf{1}\{Y_1 \leqslant \theta\} - \mathbf{1}\{f_1 \leqslant \theta\}\right]$. In Algorithm 12 we state an algorithm for prediction-powered quantile estimation; see Proposition 5.2.2 for a statement of validity.

Proposition 5.2.2 (Quantile estimation)**.** *Let θ^* be the q-quantile* (5.9)*. Then, the prediction-powered confidence set in Algorithm 12 has valid coverage:* $\liminf_{n,N\to\infty} P\left(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}\right) \geqslant 1-\alpha$.

Logistic regression. In logistic regression, the target of inference is defined by

$$\theta^* = \underset{\theta\in\mathbb{R}^d}{\arg\min}\, \mathbb{E}[\ell_\theta(X_1, Y_1)] = \underset{\theta\in\mathbb{R}^d}{\arg\min}\, \mathbb{E}\left[-Y_1\theta^\top X_1 + \log(1+\exp(\theta^\top X_1))\right], \tag{5.11}$$

where $Y_1 \in \{0,1\}$. The logistic loss is differentiable and hence the optimality condition (5.4) is ensured. Its gradient is equal to $g_\theta(x,y) = -xy + x\mu_\theta(x)$, where $\mu_\theta(x) = 1/(1+\exp(-x^\top\theta))$ is the predicted mean for point $x \in \mathcal{X}$ based on parameter vector θ. Other generalized linear models (GLMs) have the same gradient form, and thus also optimality condition (5.4), but for a different mean predictor $\mu_\theta(x)$ (see Chapter 3 of Efron [58]). For example, Poisson regression uses $\mu_\theta(x) = \exp(x^\top\theta)$. In view of our general solution for convex estimation, the rectifier is constant for all θ and equal to $\boldsymbol{\Delta}^f(\theta) \equiv \boldsymbol{\Delta}^f = \mathbb{E}\left[X_1(f_1 - Y_1)\right]$. In Algorithm 13 we state a method for prediction-powered logistic regression and in Proposition 5.2.3 we provide its guarantee. We use $X_{i,j}$ to denote the j-th coordinate of point X_i. Poisson regression is handled in essentially the same way: concretely, in Algorithm 13 we simply change the choice of $\mu_\theta(x)$ defined in line 5.

Proposition 5.2.3 (Logistic regression)**.** *Let θ^* be the logistic regression solution* (5.11)*. Then, the prediction-powered confidence set in Algorithm 13 has valid coverage:* $\liminf_{n,N\to\infty} P\left(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}\right) \geqslant 1-\alpha$.

Linear regression. Finally, we consider inference for linear regression:

$$\theta^* = \underset{\theta\in\mathbb{R}^d}{\arg\min}\, \mathbb{E}[\ell_\theta(X_1, Y_1)] = \underset{\theta\in\mathbb{R}^d}{\arg\min}\, \mathbb{E}[(Y_1 - X_1^\top\theta)^2]. \tag{5.12}$$

While it is possible to obtain an algorithm for linear regression based on Theorem 5.2.1, one can derive a more powerful solution by using the fact that the natural estimator for problem (5.12) is linear in Y. We exploit these further properties in Algorithm 14 and Proposition 5.2.4, where we state a method for prediction-powered linear regression and establish its validity, respectively.

Proposition 5.2.4 (Linear regression)**.** *Let θ^* be the linear regression solution* (5.12) *and fix $j^* \in [d]$. Then, the prediction-powered confidence interval in Algorithm 14 has valid coverage:* $\liminf_{n,N\to\infty} P\left(\theta^*_{j^*} \in \mathcal{C}_\alpha^{\mathrm{PP}}\right) \geqslant 1- \alpha$.

Algorithm 11 Prediction-powered mean estimation

```
Require: labeled data $(X, Y)$, unlabeled features $\widetilde{X}$, predictor $f$, error level $\alpha \in (0,1)$
$\hat{\theta}^{\mathrm{PP}} \leftarrow \tilde{\theta}^f - \hat{\boldsymbol{\Delta}}^f := \frac{1}{N}\sum_{i=1}^N \tilde{f}_i - \frac{1}{n}\sum_{i=1}^n (f_i - Y_i)$
$\hat{\sigma}^2_{\tilde{f}} \leftarrow \frac{1}{N}\sum_{i=1}^N (\tilde{f}_i - \tilde{\theta}^f)^2$
$\hat{\sigma}^2_{f-Y} \leftarrow \frac{1}{n}\sum_{i=1}^n (f_i - Y_i - \hat{\boldsymbol{\Delta}}^f)^2$
$w_\alpha \leftarrow z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2_{f-Y}}{n} + \frac{\hat{\sigma}^2_{\tilde{f}}}{N}}$
Output: prediction-powered confidence set $\mathcal{C}_\alpha^{\mathrm{PP}} = \left(\hat{\theta}^{\mathrm{PP}} \pm w_\alpha\right)$
```

Algorithm 12 Prediction-powered quantile estimation

```
Require: labeled data $(X, Y)$, unlabeled features $\widetilde{X}$, predictor $f$, quantile $q \in (0,1)$, error level $\alpha \in (0,1)$
Construct fine grid $\Theta_{\mathrm{grid}}$ between $\min_{i\in[N]} \tilde{f}_i$ and $\max_{i\in[N]} \tilde{f}_i$
for $\theta \in \Theta_{\mathrm{grid}}$ do
    $\hat{\boldsymbol{\Delta}}^f(\theta) \leftarrow \frac{1}{n}\sum_{i=1}^n (\mathbf{1}\{Y_i \leqslant \theta\} - \mathbf{1}\{f_i \leqslant \theta\})$
    $\hat{F}(\theta) \leftarrow \frac{1}{N}\sum_{i=1}^N \mathbf{1}\left\{\tilde{f}_i \leqslant \theta\right\}$
    $\hat{\sigma}^2_\Delta(\theta) \leftarrow \frac{1}{n}\sum_{i=1}^n \left(\mathbf{1}\{Y_i \leqslant \theta\} - \mathbf{1}\{f_i \leqslant \theta\} - \hat{\boldsymbol{\Delta}}^f(\theta)\right)^2$
    $\hat{\sigma}^2_{\tilde{f}}(\theta) \leftarrow \frac{1}{N}\sum_{i=1}^N \left(\mathbf{1}\left\{\tilde{f}_i \leqslant \theta\right\} - \hat{F}(\theta)\right)^2$
    $w_\alpha(\theta) \leftarrow z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2_\Delta(\theta)}{n} + \frac{\hat{\sigma}^2_{\tilde{f}}(\theta)}{N}}$
end for
Output: prediction-powered confidence set $\mathcal{C}_\alpha^{\mathrm{PP}} = \left\{\theta : |\hat{F}(\theta) + \hat{\boldsymbol{\Delta}}^f(\theta) - q| \leqslant w_\alpha(\theta)\right\}$
```

Algorithm 13 Prediction-powered logistic regression

```
Require: labeled data $(X, Y)$, unlabeled features $\widetilde{X}$, predictor $f$, error level $\alpha \in (0,1)$
Construct fine grid $\Theta_{\mathrm{grid}} \subset \mathbb{R}^d$ of possible coefficients
$\hat{\boldsymbol{\Delta}}^f_j \leftarrow \frac{1}{n}\sum_{i=1}^n X_{i,j}(f_i - Y_i), \quad j \in [d]$
$\hat{\sigma}^2_{\Delta,j} \leftarrow \frac{1}{n}\sum_{i=1}^n \left(X_{i,j}(f_i - Y_i) - \hat{\boldsymbol{\Delta}}^f_j\right)^2, \quad j \in [d]$
for $\theta \in \Theta_{\mathrm{grid}}$ do
    $\hat{g}^f_j(\theta) \leftarrow \frac{1}{N}\sum_{i=1}^N \widetilde{X}_{i,j}\left(\mu_\theta(\widetilde{X}_i) - \tilde{f}_i\right), \quad j \in [d], \quad$ where $\mu_\theta(x) = \frac{1}{1+\exp(-x^\top\theta)}$
    $\hat{\sigma}^2_{g,j}(\theta) \leftarrow \frac{1}{N}\sum_{i=1}^N \left(\widetilde{X}_{i,j}(\mu_\theta(\widetilde{X}_i) - \tilde{f}_i) - \hat{g}^f_j(\theta)\right)^2, \quad j \in [d]$
    $w_{\alpha,j}(\theta) \leftarrow z_{1-\alpha/(2d)}\sqrt{\frac{\hat{\sigma}^2_{\Delta,j}}{n} + \frac{\hat{\sigma}^2_{g,j}(\theta)}{N}}, \quad j \in [d]$
end for
Output: prediction-powered confidence set $\mathcal{C}_\alpha^{\mathrm{PP}} = \{\theta : |\hat{g}^f_j(\theta) + \hat{\boldsymbol{\Delta}}^f_j| \leqslant w_{\alpha,j}(\theta), \forall j \in [d]\}$
```

Algorithm 14 Prediction-powered linear regression

```
Require: labeled data $(X, Y)$, unlabeled features $\widetilde{X}$, predictor $f$, coefficient $j^* \in [d]$, error
    level $\alpha \in (0, 1)$
$\hat{\theta}^{\mathrm{PP}} \leftarrow \tilde{\theta}^f - \hat{\boldsymbol{\Delta}}^f := \widetilde{X}^\dagger \tilde{f} - X^\dagger (f - Y)$
$\tilde{\Sigma} \leftarrow \frac{1}{N} \widetilde{X}^\top \widetilde{X}$, $\tilde{M} \leftarrow \frac{1}{N} \sum_{i=1}^N (\tilde{f}_i - \widetilde{X}_i^\top \tilde{\theta}^f)^2 \widetilde{X}_i \widetilde{X}_i^\top$
$\tilde{V} \leftarrow \tilde{\Sigma}^{-1} \tilde{M} \tilde{\Sigma}^{-1}$
$\Sigma \leftarrow \frac{1}{n} X^\top X$, $M \leftarrow \frac{1}{n} \sum_{i=1}^n (f_i - Y_i - X_i^\top \hat{\boldsymbol{\Delta}}^f)^2 X_i X_i^\top$
$V \leftarrow \Sigma^{-1} M \Sigma^{-1}$
$w_\alpha \leftarrow z_{1-\alpha/2} \sqrt{\frac{V_{j^*j^*}}{n} + \frac{\tilde{V}_{j^*j^*}}{N}}$
Output: prediction-powered confidence set $\mathcal{C}_\alpha^{\mathrm{PP}} = \left( \hat{\theta}_{j^*}^{\mathrm{PP}} \pm w_\alpha \right)$
```

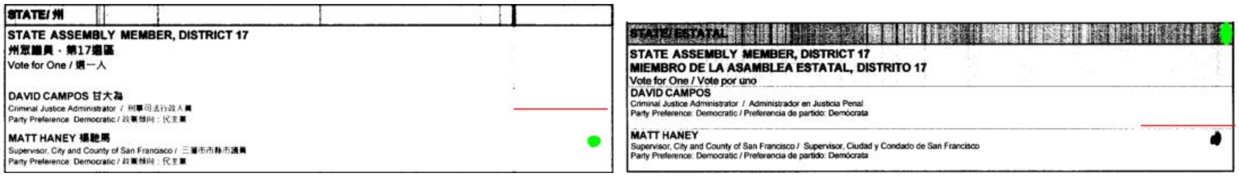
5.3 Applications

In this section we demonstrate prediction-powered inference on real tasks. In each of the following applications, we compute the prediction-powered confidence interval for an estimand of interest and compare it to two alternatives: the classical interval, which uses only the gold-standard data (X, Y), and the imputed interval, which uses only the imputed data $(\widetilde{X}, \tilde{f})$ by treating it as gold-standard data. In all cases, we show that the imputed interval, which does not account for the prediction errors, does not contain the true value of the estimand. For the two intervals that are guaranteed to be valid—prediction-powered and classical—we compare their widths as a function of n, the amount of labeled data used.

5.3.1 Auditing electronic voting

We studied audits of electronic voting in an election with two candidates. Specifically, we aimed to construct a confidence interval for the proportion of people voting for each candidate using a small number of hand-counted ballots and a large number of ballots read with an optical scanner. On Election Day in the United States, most voters use electronic or optical-scan ballots [46], neither of which are perfectly accurate. Our data were taken from a special election in San Francisco for the Assembly District 17 seat on April 19, 2022. The candidates were David Campos and Matt Haney. We constructed a prediction-powered confidence interval using an optical ballot labeling system and a small number of ballots which we labeled ourselves. This is an example of a *risk-limiting audit*—a statistically valid way to check the results of an election by inspecting subsets of ballots [see, e.g., 112].

Formally, we have $N = 78150$ images of paper ballots, $\widetilde{X}_i \in \mathcal{X}$, $i \in [N]$, taken using an optical ballot scanner. Each ballot has an associated ground-truth binary vote, $\widetilde{Y}_i \in \{0, 1\}$, $i \in [N]$, where a "1" indicates a vote for Matt Haney and a "0" indicates a vote for David Campos. The target of inference is the fraction of votes for Matt Haney, $\theta^* = \mathbb{E}[\widetilde{Y}_1]$. To compute the intervals, we hand-annotated $n = 1024$ randomly sampled ballots and imputed



STATE/ 州

STATE ASSEMBLY MEMBER, DISTRICT 17
州眾議員 · 第17選區
Vote for One / 選一人

DAVID CAMPOS 甘大為
Criminal Justice Administrator / 刑事司法行政人員
Party Preference: Democratic / 政黨傾向：民主黨

MATT HANEY 楊馳馬
Supervisor, City and County of San Francisco / 三藩市市縣市議員
Party Preference: Democratic / 政黨傾向：民主黨

STATE/ ESTATAL

STATE ASSEMBLY MEMBER, DISTRICT 17
MIEMBRO DE LA ASAMBLEA ESTATAL, DISTRITO 17
Vote for One / Vote por uno

DAVID CAMPOS
Criminal Justice Administrator / Administrador en Justicia Penal
Party Preference: Democratic / Preferencia de partido: Demócrata

MATT HANEY
Supervisor, City and County of San Francisco / Supervisor, Ciudad y Condado de San Francisco
Party Preference: Democratic / Preferencia de partido: Demócrata

Figure 5.1: **Examples of ballots** correctly and incorrectly classified. The raw ballot is black and white, the voter's marking is automatically identified by a computer vision algorithm with a green annotation, and markings below the red line annotation will be considered votes for Matt Haney (and vice versa). The instructional portion of the ballots was cropped out.

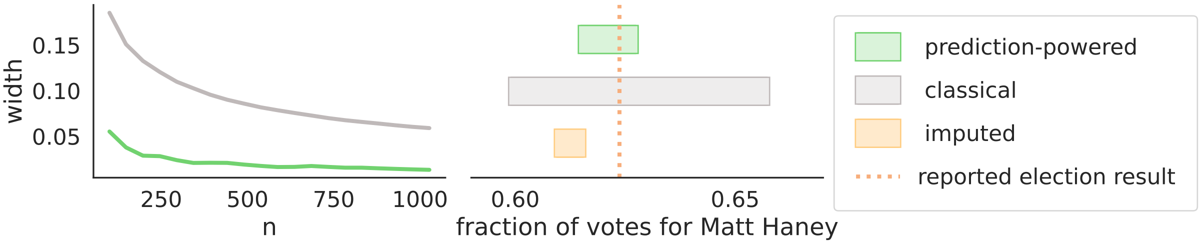


Figure 5.2: **Election results** produced by prediction-powered inference and the classical and imputed baselines at level 95%. Left: width of intervals as a function of n. Right: confidence intervals with $n = 1024$.

labels using a computer vision algorithm $f : \mathcal{X} \to \{0, 1\}$, representing an optical-ballot scanner. The accuracy of f is 99%, and it is biased towards Matt Haney due to printing errors in the ballots; see Figure 5.1 for examples. We used Algorithm 11 to construct the prediction-powered confidence interval, and binomial confidence intervals for the imputed and classical baselines. The prediction-powered interval has roughly 1/4 the width of its classical counterpart, and the interval based on imputation is invalid—it does not cover the ground truth. See Figure 5.2.

5.3.2 Relating protein structure and post-translational modifications

We demonstrate how prediction-powered confidence intervals for the mean can be used to construct confidence intervals for more elaborate estimands, such as the odds ratio, which is commonly used to quantify associations between binary random variables.

The goal in this section is to characterize the structural context of post-translational modifications (PTMs), which are biochemical modifications of specific positions of a protein

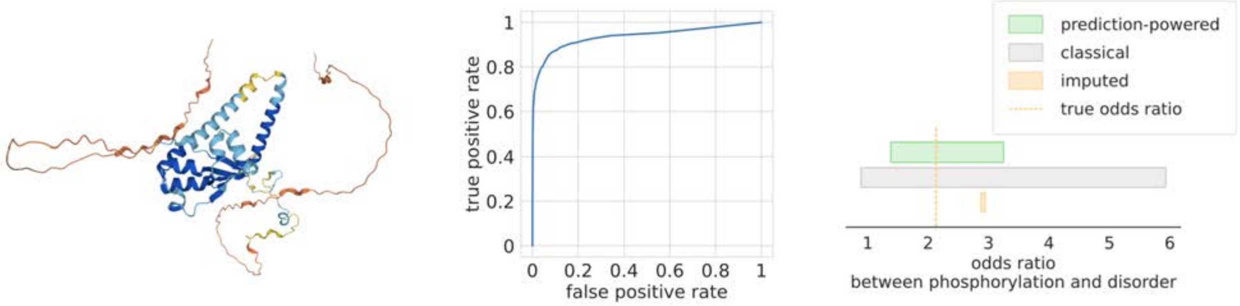


Figure 5.3: **AlphaFold-based prediction of disorder**. Left: predicted disorder for one example protein (UniProt S5FZ81), colored by predicted probability of disorder per position. Middle: ROC curve of disorder prediction based on AlphaFold structure. Right: confidence interval for the odds ratio between disorder and phosphorylation (type of PTM) produced by prediction-powered inference and the classical and imputed baselines, when $n = 571$. Unlike the classical interval, the prediction-powered interval excludes the value of one and thus the direction of the association is unambiguous.

sequence that play important regulatory roles. One question of interest is whether PTMs occur more frequently in particular contexts within a protein's three-dimensional structure, such as intrinsically disordered regions (IDRs), segments of a protein that do not abide in a fixed three-dimensional structure. Recently, Bludau et al. [16] studied this relationship on an unprecedented proteome-wide scale by using AlphaFold-predicted structures [92] to predict IDRs, in contrast to previous work which considered far fewer experimentally derived structures.

We will refer to a position of a protein sequence being/not being in an IDR as "disordered"/"ordered," and having/not having a PTM as "modified"/"unmodified." Let $Y_i \in \{1, 0\}$ denote the gold-standard label of whether or not a position is disordered, and let $Z_i \in \{1, 0\}$ denote whether or not a position is modified. Following Bludau et al. [16], we obtain a prediction for Y_i, denoted $f_i \in \{1, 0\}$, based on the protein structure predicted by AlphaFold. To quantify the association between PTMs and IDRs, the authors computed the odds ratio between f_i and Z_i on a data set of hundreds of thousands of protein sequence positions. Though some of the data points also contained a gold-standard label, Y_i, Bludau et al. [16] did not use these labels in their analyses to avoid dealing with conflicts between labels and predictions. Here, we show how to use both labels and predictions to give confidence intervals for the odds ratio that are valid, in contrast to intervals based only on f_i, and smaller than intervals obtained only using Y_i.

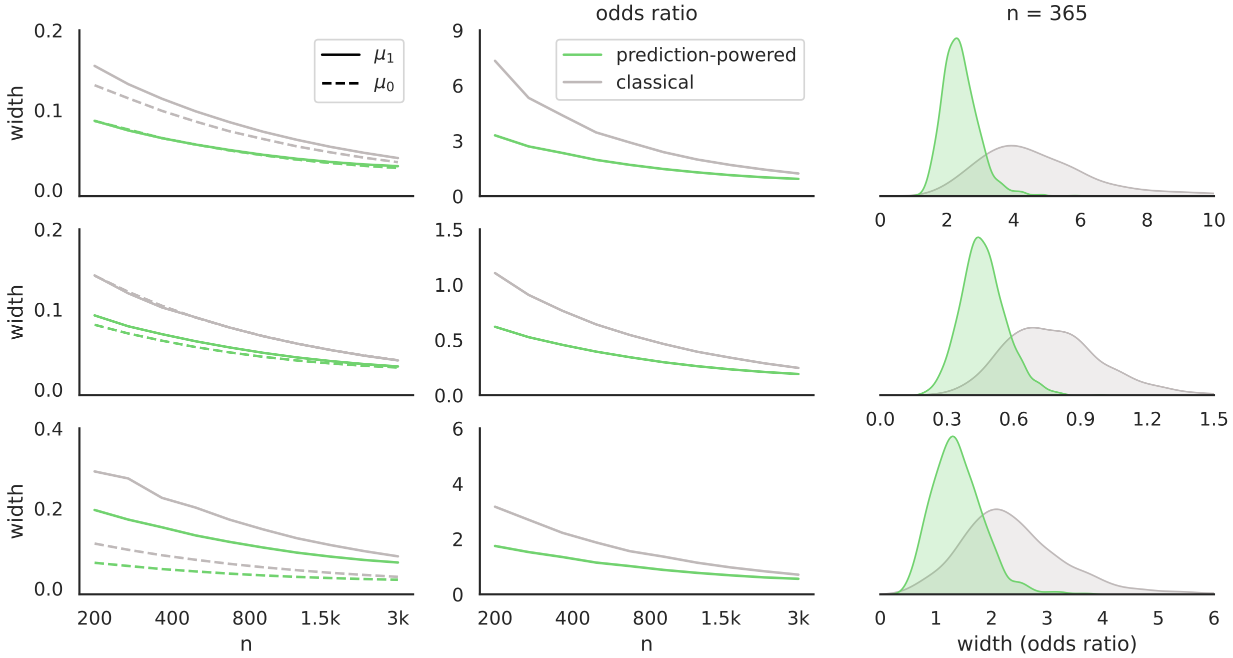


Figure 5.4: **Odds ratio for three different PTMs**, phosphorylation (top row), ubiquitination (middle row), and acetylation (bottom row). Left: widths of prediction-powered and classical confidence intervals for μ_1 (solid line) and μ_0 (dashed line). Middle: widths of prediction-powered and classical confidence intervals for the odds ratio. Right: distribution of interval widths for the odds ratio when $n = 365$.

The odds ratio between Y_i and Z_i can be written as a function of two means:

$$\theta^* = \frac{\mu_1/(1-\mu_1)}{\mu_0/(1-\mu_0)}, \tag{5.13}$$

where $\mu_1 = P(Y = 1 \mid Z = 1)$ and $\mu_0 = P(Y = 1 \mid Z = 0)$. We therefore proceed by constructing $1 - \alpha/2$ prediction-powered confidence intervals for μ_0 and μ_1, denoted $\mathcal{C}_0^{\mathrm{PP}} = [l_0, u_0]$ and $\mathcal{C}_1^{\mathrm{PP}} = [l_1, u_1]$, respectively. We then propagate $\mathcal{C}_0^{\mathrm{PP}}$ and $\mathcal{C}_1^{\mathrm{PP}}$ through the odds-ratio formula (5.13) to get the following confidence interval:

$$\mathcal{C}^{\mathrm{PP}} = \left\{ \frac{c_1}{1-c_1} \cdot \frac{1-c_0}{c_0} : c_0 \in \mathcal{C}_0^{\mathrm{PP}}, c_1 \in \mathcal{C}_1^{\mathrm{PP}} \right\} = \left(\frac{l_1}{1-l_1} \cdot \frac{1-u_0}{u_0}, \frac{u_1}{1-u_1} \cdot \frac{1-l_0}{l_0} \right). \tag{5.14}$$

By a union bound, $\mathcal{C}^{\mathrm{PP}}$ contains θ^* with probability at least $1 - \alpha$. We set $\alpha = 0.1$.

We have 10803 data points from [16], from which we simulated labeled and unlabeled data sets as follows. For each of 1000 trials, we randomly sampled n points to serve as the labeled data set and used the remaining $N = 10803 - n$ points as the unlabeled data set, where we do not observe the labels. For all values of n and all three different types of PTMs that we examined, the prediction-powered confidence intervals are smaller than classical intervals; see Figure 5.4. Often, the classical intervals are large enough that they contain the odds ratio value of one, as demonstrated in Figure 5.3, which means the direction of the association cannot be determined from the confidence interval. On the other hand, the imputed confidence interval is far too small and significantly overestimates the true odds ratio; see Figure 5.3.

5.3.3 Relationship between age, sex, and income

We next used census data to investigate the quantitative effects of age and sex on income by constructing confidence intervals for the linear regression coefficients relating age (1-99) and sex (M/F) to income. As can be seen in Figure 5.5, which plots the distribution of income across sex and different age ranges, income generally increases with age, and men earn more than women in every age category.

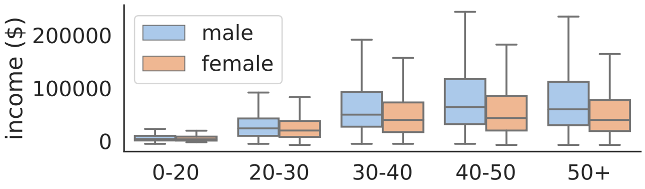


Figure 5.5: **Distribution of income stratified by age and sex** in the census year 2019.

Concretely, we used the Folktables interface [48] to download census data from California in the year 2018 (377575 people), including yearly income ($), and ten covariates including age and sex. On this data, we trained an XGBoost model [36] to predict yearly income from the covariates. Then, in the year 2019 ($N = 378817$ people), we observed all the covariates but only a small number $n = 100$ of yearly incomes. Using the model, we imputed the remaining yearly incomes from the covariates, and then regressed age and sex to the imputed yearly incomes using ordinary least squares. We then formed a prediction-powered confidence interval for the regression parameters using Algorithm 14. The classical and imputed baselines used standard least-squares confidence intervals. See Figure 5.6 for results. Note that income is difficult to predict, as evidenced by the width of the boxplots in Figure 5.5, so the prediction-powered interval provides only a moderate improvement over the classical interval. Critically, however, it yields this improvement without succumbing to the overconfidence exhibited by the imputed interval.

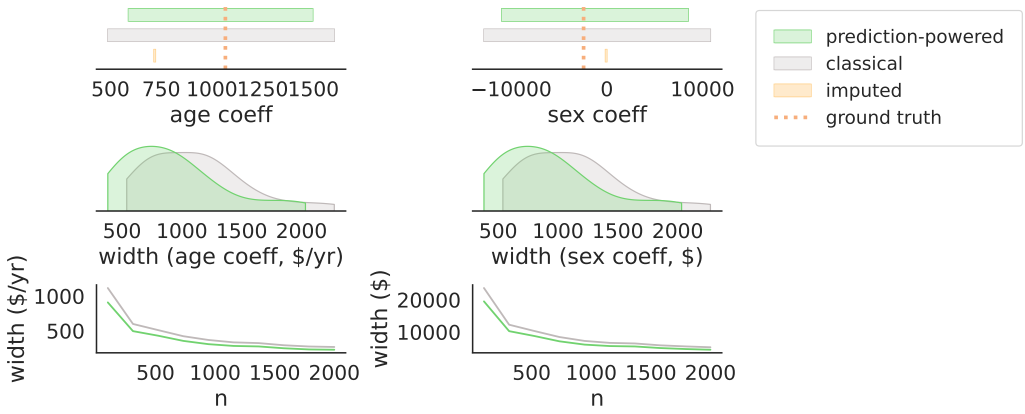


Figure 5.6: **Confidence intervals at the 95% level** are smaller using prediction-powered inference. Top: confidence intervals with $n = 100$. Middle: density of interval width with $n = 100$. Bottom: average width as a function of n.

5.3.4 Relationship between income and private health insurance

Again using census data, we studied the effect of income on the procurement of private health insurance. In particular, we fit a logistic regression relating an individual's income to the probability they have private health insurance. Generally, the higher a person's income, the more likely they are to have private health insurance.

The setup is essentially the same as in Section 5.3.3: we used the Folktables interface to download California census data on income, a binary indicator of private health insurance, and other predictive covariates. We used XGBoost to impute the predictions, and Algorithm 13 to construct the intervals; see Figure 5.7 for results.

5.3.5 Distribution of gene expression levels

In this section, we demonstrate the construction of prediction-powered confidence intervals on quantiles for studying the effects of regulatory DNA on gene expression. In particular, we aim to characterize the distribution of gene expression levels induced by a population of promoters—regulatory DNA sequences that control how frequently a gene is transcribed. Recently, Vaishnav et al. [166] trained a state-of-the-art transformer model on tens of millions of random promoter sequences with the goal of predicting the expression level of a particular gene induced by a promoter sequence (see Figure 5.8). They then used the model's predictions to study the effects of promoters—for example, by assessing how quantiles of predicted expression levels differ between different populations of promoters, and verifying those observations by experimentally measuring the expression levels of the promoters of interest.

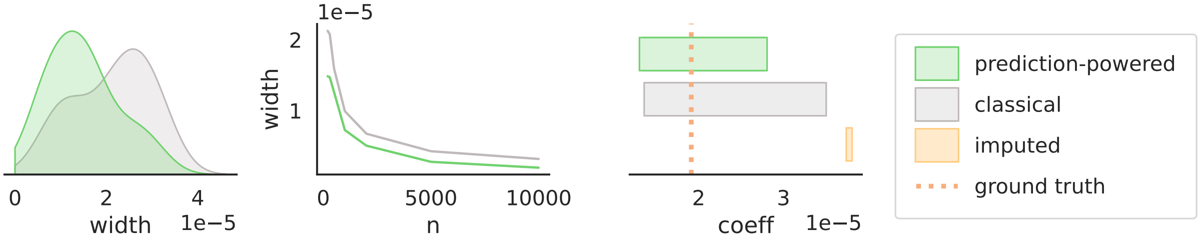


Figure 5.7: **Confidence intervals for the logistic regression coefficient** relating income and private health insurance coverage at the 95% level. Left: distribution of interval widths with $n = 200$. Middle: mean width as a function of n. Right: intervals with $n = 200$.

Let X_i be an 80-base-pair promoter sequence for a particular gene, and let $Y_i \in [0, 20]$ denote a measurement of the expression level it causes for the gene. Furthermore, let $f_i \in [0, 20]$ denote the corresponding expression level predicted by the transformer model in [166]. We focus on estimating the 0.25-, 0.5-, and 0.75-quantiles of expression levels induced by native yeast promoters—promoter sequences that are naturally found in the genomes of *S. cerevisiae*.

We have 61150 labeled native yeast promoter sequences from [166], from which we simulated labeled and unlabeled data sets as follows. For each of 1000 trials, we randomly

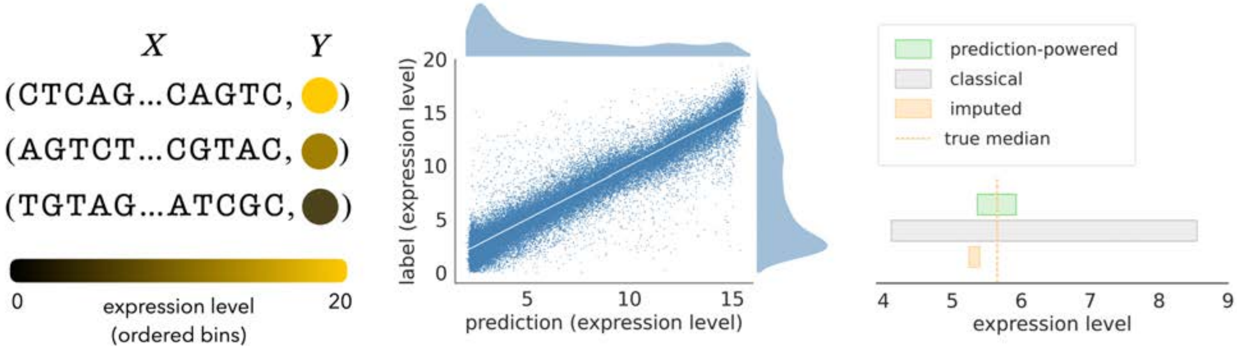


Figure 5.8: **Predicting gene expression levels from a promoter sequence** [166]. Left: each data point consists of a promoter sequence, X_i, and an expression level, Y_i. Middle: predictive performance of the transformer model on the native yeast promoters used in our experiments (RMSE 2.18, Pearson 0.963, Spearman 0.946). Right: confidence intervals for the median native yeast promoter expression level with $n = 75$ and $\alpha = 0.1$.

sampled n points to serve as the labeled data set and used the remaining $N = 61150 - n$ points as the unlabeled data set. We then used Algorithm 12 to construct prediction-powered intervals with $\alpha = 0.1$. The prediction-powered confidence intervals for all three quantiles are much smaller than the classical intervals for all values of n; see Figure 5.9.

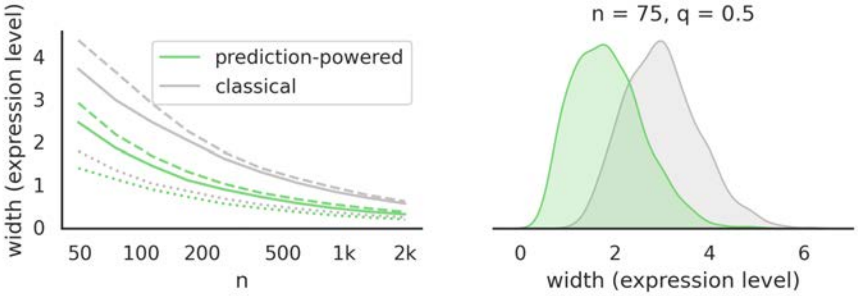


Figure 5.9: **Widths of confidence intervals around the median** using the transformer model of expression level developed by Vaishnav et al. [166]. Left: average width of prediction-powered and classical confidence intervals for the 0.25-quantile (dashed lines), 0.5-quantile (solid lines), and 0.75-quantile (dotted lines). Right: distribution of confidence interval widths for the median using $n = 75$.

5.3.6 Counting plankton

We counted the number of plankton observed by the Imaging FlowCytobot [131, 132], an automated, submersible flow cytometry system, at Woods Hole Oceanographic Institution in the year 2014. We also had access to data from previous years, which we treated as labeled, and the 2014 data were fully imputed. The resulting confidence interval does not assume that the 2014 data are identically distributed to the data from previous years; we explicitly adjust for the fact that the probability of observing a plankton may change using the label shift technique from Section 5.4.2.

More formally, our inputs $\widetilde{X}_i$ are images taken by the flow cytometry system and the labels $\widetilde{Y}_i$ are one of $\{\text{detritus}, \text{plankton}\}$, where detritus represents unspecified organic matter; see Figure 5.10 for examples. We are interested in the number of plankton in the year 2014. For the machine-learning algorithm, we fine-tune an ImageNet pretrained ResNet-152 [77] on labeled data from the years 2006-2012 (2812527 data points). The labeled data set consists of $n = 421238$ image–label pairs from 2013 that the model was not trained on. Finally, we received $N = 329832$ unlabeled images $\widetilde{X}_i$ from the new year that has undergone a label shift. Given these three ingredients, we used the technique in Section 5.4.2 to construct the prediction-powered confidence interval on the frequency of observed plankton, while accounting for the distribution shift. We then propagated the confidence interval into a count. See Figure 5.11 for results. Note that assuming the data is i.i.d. and applying

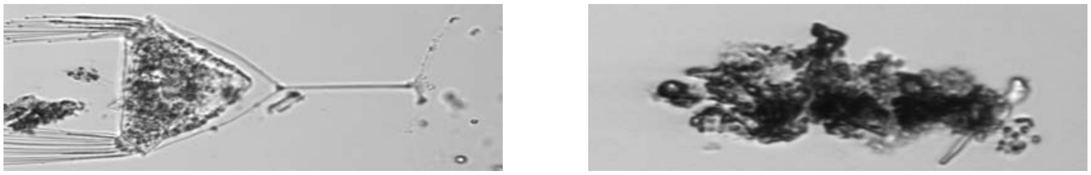


Figure 5.10: **Examples of plankton and detritus**, respectively.

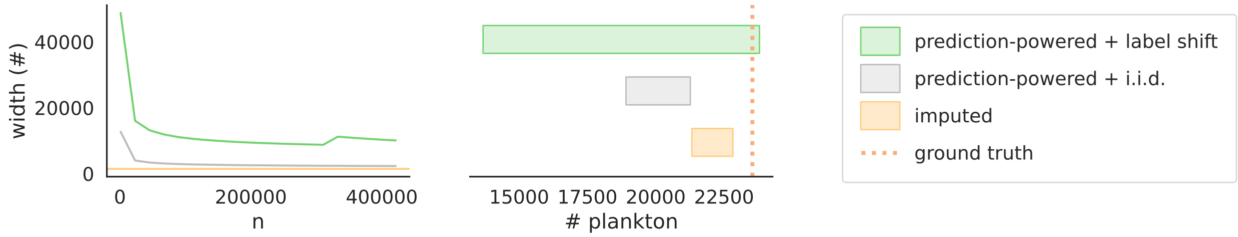


Figure 5.11: **Confidence intervals on the number of plankton** observed in 2014 at the 95% level. Left: mean width as a function of n. Right: confidence intervals with $n = 421238$.

Algorithm 11 for mean estimation results in biased inferences, as does the imputed baseline using a classical binomial confidence interval.

5.4 Extensions

We demonstrate that the framework of prediction-powered inference is applicable beyond the setting of i.i.d. convex estimation studied in Section 5.2. First, we provide a strategy for prediction-powered inference when θ^* can be expressed as the optimum of any optimization problem, not necessarily a convex one. Then, we discuss prediction-powered inference under certain forms of distribution shift.

5.4.1 Beyond convex estimation

The tools developed in Section 5.2 were tailored to unconstrained convex optimization problems. In general, however, inferential targets can be defined in terms of nonconvex losses or they may have (possibly even nonconvex) constraints. For such general optimization problems, we cannot expect the condition (5.4) to hold. In this section we generalize our approach to a broad class of risk minimizers:

$$\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}[\ell_\theta(X_1, Y_1)], \tag{5.15}$$

where $\ell_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a possibly nonconvex loss function and Θ is an arbitrary set of admissible parameters. As before, if θ^* is not a unique minimizer, our method will return a set that contains all minimizers.

The problem (5.15) subsumes all previously studied settings. Indeed, when the loss ℓ_θ is convex and subdifferentiable and $\Theta = \mathbb{R}^p$ for some p—which is the case for all problems previously studied—θ^* can be equivalently characterized via the condition (5.4). In this section we provide a solution that can handle problems of the form (5.15) in full generality. We note, however, that the solution does not reduce to the one in Section 5.2 for convex estimation problems, and we expect the method from Section 5.2 to be more powerful for convex estimation problems with low-dimensional rectifiers.

To correct the imputation approach, we rely on the following rectifier:

$$\boldsymbol{\Delta}^f(\theta) = \mathbb{E}\left[\ell_\theta(X_1, Y_1) - \ell_\theta(X_1, f_1)\right]. \tag{5.16}$$

Notice that the rectifier (5.16) is always one-dimensional, while the rectifier (5.5) was p-dimensional.

One key difference relative to the approach of Section 5.2 is that we have an additional step of data splitting. We need the additional step because, unlike in convex estimation where we know $\mathbb{E}[g_{\theta^*}(X_1, Y_1)] = 0$, for general problems we do not know the value of $\mathbb{E}[\ell_{\theta^*}(X_1, Y_1)]$. To circumvent this issue, we estimate $\mathbb{E}[\ell_{\theta^*}(X_1, Y_1)]$ by approximating θ^* with an imputed estimate on the first $N/2$ unlabeled data points (for simplicity, take N to be even). To state the main result, we define

$$\tilde{\theta}^f = \arg\min_{\theta\in\Theta} \frac{2}{N}\sum_{i=1}^{N/2} \ell_\theta(\tilde{X}_i, \tilde{f}_i), \quad \tilde{L}^f(\theta) := \frac{2}{N}\sum_{i=N/2+1}^{N} \ell_\theta(\tilde{X}_i, \tilde{f}_i).$$

Theorem 5.4.1 (General risk minimization). *Fix $\alpha \in (0,1)$ and $\delta \in (0,\alpha)$. Suppose that, for any $\theta \in \Theta$, we can construct $\left(\mathcal{R}^l_{\delta/2}(\theta), \mathcal{R}^u_{\delta/2}(\theta)\right)$ and $\left(\mathcal{T}^l_{\frac{\alpha-\delta}{2}}(\theta), \mathcal{T}^u_{\frac{\alpha-\delta}{2}}(\theta)\right)$ such that*

$$\begin{aligned}
&P\left(\boldsymbol{\Delta}^f(\theta) \leqslant \mathcal{R}^u_{\delta/2}(\theta)\right) \geqslant 1-\delta/2; \quad P\left(\boldsymbol{\Delta}^f(\theta) \geqslant \mathcal{R}^l_{\delta/2}(\theta)\right) \geqslant 1-\delta/2;\\
&P\left(\tilde{L}^f(\theta) - \mathbb{E}[\ell_\theta(X_1, f_1)] \leqslant \mathcal{T}^u_{\frac{\alpha-\delta}{2}}(\theta)\right) \geqslant 1 - \frac{\alpha-\delta}{2};\\
&P\left(\tilde{L}^f(\theta) - \mathbb{E}[\ell_\theta(X_1, f_1)] \geqslant \mathcal{T}^l_{\frac{\alpha-\delta}{2}}(\theta)\right) \geqslant 1 - \frac{\alpha-\delta}{2}.
\end{aligned}$$

Let

$$\mathcal{C}^{\mathrm{PP}}_\alpha = \left\{\theta \in \Theta : \tilde{L}^f(\theta) \leqslant \tilde{L}^f(\tilde{\theta}^f) - \mathcal{R}^l_{\delta/2}(\theta) + \mathcal{R}^u_{\delta/2}(\tilde{\theta}^f) + \mathcal{T}^u_{\frac{\alpha-\delta}{2}}(\theta) - \mathcal{T}^l_{\frac{\alpha-\delta}{2}}(\tilde{\theta}^f)\right\}.$$

Then, we have

$$P\left(\theta^* \in \mathcal{C}^{\mathrm{PP}}_\alpha\right) \geqslant 1-\alpha.$$

For example, if the loss $\ell_\theta(x,y)$ takes values in $[0,B]$ for all x,y, then we can set $\mathcal{T}_{\alpha-\delta}(\theta) = B\sqrt{\frac{\log(1/(\alpha-\delta))}{N}}$. The validity of this choice follows by Hoeffding's inequality.

Mode estimation. A commonplace inference task that does not fall under convex estimation is the problem of estimating the mode of the outcome distribution. When the outcome takes values in a discrete set Θ, this can be done by using the loss function $\ell_\theta(y) = \mathbf{1}\{y \neq \theta\}, \theta \in \Theta$. A generalization of this approach to continuous outcome distributions is obtained by defining the loss $\ell_\theta(y) = \mathbf{1}\{|y - \theta| > \eta\}$, for some width parameter $\eta > 0$. The target of inference is thus the point $\theta \in \mathbb{R}$ that has the most probability mass in its η-neighborhood, $\theta^* = \arg\min_{\theta\in\mathbb{R}} P(|Y_1 - \theta| > \eta)$. Theorem 5.4.1 applies directly in both the discrete and continuous cases.

Tukey's biweight robust mean. The Tukey biweight loss function is a commonly used loss in robust statistics that results in an outlier-robust mean estimate. It behaves approximately like a quadratic near the origin and is constant far away from the origin. Formally, Tukey's biweight loss function is given by

$$\ell_\theta(y) = \begin{cases} \frac{c^2}{6}\left(1 - \left(1 - \frac{(y-\theta)^2}{c^2}\right)^3\right), & |y - \theta| \leqslant c, \\ \frac{c^2}{6}, & \text{otherwise,} \end{cases}$$

where c is a user-specified tuning parameter. It is not hard to see that the function $\ell_\theta(y)$ is nonconvex and hence not amenable to the analysis in Section 5.2; however, Theorem 5.4.1 applies.

Model selection. Nonconvex risk minimization problems are ubiquitous in model selection. For example, a common model selection strategy is best subset selection, which optimizes the squared loss, $\ell_\theta(x, y) = (y - x^\top\theta)^2$, subject to the constraint $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_0 \leqslant k\}$. Here, Θ is the space of all k-sparse vectors for a user-chosen parameter k. Even though the loss function is convex, Θ is a nonconvex constraint set and hence we cannot rely on the condition (5.4) to find the minimizer. However, Theorem 5.4.1 still applies.

5.4.2 Inference under distribution shift

In Section 5.2 we focused on forming prediction-powered confidence intervals when the labeled and unlabeled data come from the same distribution. Herein, we extend our tools to the case where the labeled data (X, Y) comes from $\mathbb{P}$ and the unlabeled data $(\widetilde{X}, \widetilde{Y})$—which defines the target of inference θ^*—comes from $\mathbb{Q}$, and these are related by either a label shift or a covariate shift. For covariate shift, we handle all estimation problems previously studied; for label shift, we handle certain types of linear problems.

We will write $\mathbb{E}_{\mathbb{Q}}, \mathbb{E}_{\mathbb{P}}$, etc to indicate which distribution the data inside the expectation is sampled from.

Covariate shift

First, we assume that $\mathbb{Q}$ is a known *covariate shift* of $\mathbb{P}$. That is, if we denote by $\mathbb{Q} = \mathbb{Q}_X \cdot \mathbb{Q}_{Y|X}$ and $\mathbb{P} = \mathbb{P}_X \cdot \mathbb{P}_{Y|X}$ the relevant marginal and conditional distributions, we assume that $\mathbb{Q}_{Y|X} = \mathbb{P}_{Y|X}$. As in previous sections, we consider estimands of the form

$$\theta^* = \arg\min_{\theta \in \Theta} \mathop{\mathbb{E}}_{\mathbb{Q}}[\ell_\theta(X_1, Y_1)]. \tag{5.17}$$

Estimands of the form (5.17) can be related to risk minimizers on $\mathbb{P}$ using the Radon-Nikodym derivative. In particular, suppose that $\mathbb{Q}_X$ is dominated by $\mathbb{P}_X$ and assume that the Radon-Nikodym derivative $w(x) = \frac{\mathbb{Q}_X}{\mathbb{P}_X}(x)$ is known. Then, we can rewrite (5.17) as

$$\theta^* = \arg\min_{\theta \in \Theta} \mathop{\mathbb{E}}_{\mathbb{P}}[\ell_\theta^w(X_1, Y_1)],$$

where $\ell_\theta^w(x, y) = w(x)\ell_\theta(x, y)$. In words, risk minimizers on $\mathbb{Q}$ can simply be written as risk minimizers on $\mathbb{P}$, but with a reweighted loss function. This permits inference on the rectifier to be based on data sampled from $\mathbb{P}$ as before. For concreteness, we explain the approach in detail for convex risk minimizers. Let

$$\mathbf{\Delta}^{f,w}(\theta) = \mathop{\mathbb{E}}_{\mathbb{P}}\left[g_\theta^w(X_1, Y_1) - g_\theta^w(X_1, f_1)\right], \tag{5.18}$$

where $g_\theta^w(x, y) = g_\theta(x, y) \cdot w(x)$ and g_θ is a subgradient of ℓ_θ as before. A confidence set for the above rectifier suffices for prediction-powered inference on θ^*.

Corollary 5.4.1 (Covariate shift)**.** *Suppose that the problem* (5.17) *is a nondegenerate convex estimation problem. Fix* $\alpha \in (0, 1)$ *and* $\delta \in (0, \alpha)$*. Suppose that, for any* $\theta \in \mathbb{R}^p$*, we can construct* $\mathcal{R}_\delta(\theta)$ *and* $\mathcal{T}_{\alpha-\delta}(\theta)$ *satisfying*

$$P\left(\mathbf{\Delta}^{f,w}(\theta) \in \mathcal{R}_\delta(\theta)\right) \geqslant 1 - \delta; \quad P\left(\mathbb{E}[g_\theta^w(X_1, f_1)] \in \mathcal{T}_{\alpha-\delta}(\theta)\right) \geqslant 1 - (\alpha - \delta).$$

Let $\mathcal{C}_\alpha^{\mathrm{PP}} = \{\theta : 0 \in \mathcal{R}_\delta(\theta) + \mathcal{T}_{\alpha-\delta}(\theta)\}$*, where + denotes the Minkowski sum. Then,*

$$P(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}) \geqslant 1 - \alpha. \tag{5.19}$$

The same reweighting principle can be used to handle nonconvex risk minimizers as in Section 5.4.1.

Label shift

Next, we analyze classification problems where the proportions of the classes in the labeled data is different from those in the unlabeled data. This problem has been studied before in the literature on domain adaptation, e.g. by Lipton et al. [113], but our treatment focuses on the formation of confidence intervals. Formally, let $\mathcal{Y} = \{1, ..., K\}$ be the label space and assume that $\mathbb{Q}_{X|Y} = \mathbb{P}_{X|Y}$. We consider estimands of the form

$$\theta^* = \mathop{\mathbb{E}}_{\mathbb{Q}_Y}[\nu(Y)],$$

where $\nu : \mathcal{Y} \to \mathbb{R}$ is a fixed function. For example, choosing $\nu(y) = \mathbf{1}\{y = k\}$ for some $k \in [K]$ asks for inference on the proportion of instances that belong to class k.

Using an analogous decomposition to the one for mean estimation, we can write

$$\theta^* = \mathop{\mathbb{E}}_{\mathbb{Q}_f}[\nu(f)] + (\mathop{\mathbb{E}}_{\mathbb{Q}_Y}[\nu(Y)] - \mathop{\mathbb{E}}_{\mathbb{Q}_f}[\nu(f)]) = \theta^f + \boldsymbol{\Delta}^f,$$

where $\mathbb{Q}_f$ denotes the distribution of $f(X), X \sim \mathbb{Q}_X$. The quantity θ^f can be estimated using the unlabeled data from $\mathbb{Q}$ and the model. Estimating the quantity $\boldsymbol{\Delta}^f$ using samples from $\mathbb{P}$ will require leveraging the structure of the distribution shift. Central to our analysis will be the confusion matrix

$$\mathcal{K}_{j,l} = \mathbb{Q}\left(f(X) = j \;\middle|\; Y = l\right), \; j, l \in [K]. \tag{5.20}$$

The label-shift assumption implies that $\mathcal{K}_{j,l} = \mathbb{P}\left(f(X) = j \mid Y = l\right)$, which can be estimated from labeled data sampled from $\mathbb{P}$. In particular, we estimate $\mathcal{K}$ from the labeled data as

$$\widehat{\mathcal{K}}_{j,l} = \frac{1}{n(l)} \sum_{i=1}^{n} \mathbf{1}\{f_i = j, Y_i = l\}, \text{ where } n(l) = \sum_{i=1}^{n} \mathbf{1}\{Y_i = l\}. \tag{5.21}$$

Similarly, we can estimate $\mathbb{Q}_f(k), k \in [K]$ as

$$\widehat{\mathbb{Q}}_f(k) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left\{\tilde{f}_i = k\right\}.$$

Treating $\mathbb{Q}_f$ and $\mathbb{Q}_Y$ as vectors, notice that we can write $\mathbb{Q}_f = \mathcal{K}\mathbb{Q}_Y$, and hence $\mathbb{Q}_Y = \mathcal{K}^{-1}\mathbb{Q}_f$. This leads to a natural estimate of $\mathbb{Q}_Y$, $\widehat{\mathbb{Q}}_Y = \widehat{\mathcal{K}}^{-1}\widehat{\mathbb{Q}}_f$. Below, we use these quantities to construct a prediction-powered confidence interval for $\theta^* = \mathbb{E}_{\mathbb{Q}_Y}[\nu(Y)]$.

Theorem 5.4.2 (Label shift)**.** *Fix $\alpha \in (0, 1)$ and $\delta \in (0, \alpha)$. Let*

$$\mathcal{C}_\alpha^{\mathrm{PP}} = \left(\mathop{\mathbb{E}}_{\widehat{\mathbb{Q}}_Y}[\nu(Y)] \pm \left(\max_{l,k \in [K]} \max_{p \in C_{l,k}} |\widehat{\mathcal{K}}_{l,k} - p| + \sqrt{\frac{1}{2N} \log \frac{2}{\alpha - \delta}}\right)\right),$$

where

$$C_{l,k} = \left\{p : n(k)\widehat{\mathcal{K}}_{l,k} \in \left[F^{-1}_{\mathrm{Binom}(n(k),p)}\left(\frac{\delta}{2K^2}\right), F^{-1}_{\mathrm{Binom}(n(k),p)}\left(1 - \frac{\delta}{2K^2}\right)\right]\right\}$$

and $F_{\mathrm{Binom}(n(k),p)}$ denotes the Binomial CDF. Then,

$$P(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}) \geqslant 1 - \alpha.$$

Naturally, the confidence interval becomes more conservative as the number of classes grows. Also, the power of the bound depends on the smallest number of instances observed for a particular class.

5.5 Related work

Our technical results generalize tools from the model-assisted survey sampling literature [e.g., 148], which provides methods to improve inference from surveys in the presence of auxiliary information. In particular, the mean estimator in Section 5.1.2 is the difference estimator, closely related to generalized regression estimators [29]. It has long been recognized that model predictions can be leveraged as auxiliary data [186], and much work has gone into producing asymptotically valid confidence intervals when the predictive model is fit on the same data that is used for inference—see [18] for a recent overview. Our work is also related to the statistical literature on missing data and multiple imputation [e.g., 114]. In particular, Robins et al. [146], Robins and Rotnitzky [145], Chen and Breslow [33], Yu and Nan [187] study regression with missing data. Likewise, our setting is related to measurement error [e.g., 28], particularly to Chen et al. [37], who study the estimation of parameters defined as solutions to many estimating equations, as we will in this work.

Recently, a body of work on estimation with many labeled data points and few unlabeled data points has been developed [4, 35, 136, 176], focusing on efficiency in semiparametric or high-dimensional regimes. In particular, Chakrabortty and Cai [30] study efficient estimation of linear regression parameters, Chakrabortty et al. [31] study efficient quantile estimation, Zhang and Bradic [188] study mean estimation in a high-dimensional setting, and Hou et al. [83] study an imputation approach to improving generalized linear models. Our work continues in this vein but focuses on the setting where we have access to a good predictive model fit on separate data. This allows us to tackle a much wider range of estimands (e.g., minimizers of any convex objective) and give finite-sample inferences without assumptions about the machine-learning model. Secondly, we go beyond random sampling and consider certain forms of distribution shift in this work.

More distantly, our setting, in which we have access to some labeled data alongside unlabeled data, also appears in semisupervised learning [e.g., 190, 191]—that literature studies the question of how to improve prediction accuracy with unlabeled data. We also refer the reader to the related literature about surrogates in causal inference [e.g., 94]. Thematically, our work is most similar to the work of Wang et al. [175], who also introduce a method to correct machine-learning predictions for the purpose of subsequent inference. However, our work provides confidence intervals that are provably valid under minimal assumptions about the data-generating distribution, whereas Wang et al. require certain parametric assumptions about the relationship between the prediction model and the true response.

5.6 Deferred proofs

5.6.1 Proof of Theorem 5.2.1

We show that $\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}$ with probability at least $1-\alpha$; that is, with probability at least $1-\alpha$ it holds that

$$0 \in \mathcal{R}_\delta(\theta^*) + \mathcal{T}_{\alpha-\delta}(\theta^*).$$

Consider the event $E = \{\boldsymbol{\Delta}^f(\theta^*) \in \mathcal{R}_\delta(\theta^*)\} \cap \{\mathbb{E}[g_{\theta^*}(X_1, f_1)] \in \mathcal{T}_{\alpha-\delta}(\theta^*)\}$. By a union bound, $P(E) \geqslant 1-\alpha$. On the event E, we have that

$$\begin{aligned}\mathbb{E}[g_{\theta^*}(X_1,Y_1)] &= \mathbb{E}[g_{\theta^*}(X_1,Y_1)] - \mathbb{E}[g_{\theta^*}(X_1,f_1)] + \mathbb{E}[g_{\theta^*}(X_1,f_1)] && (5.22)\\ &= \boldsymbol{\Delta}^f(\theta^*) + \mathbb{E}[g_{\theta^*}(X_1,f_1)] \in \mathcal{R}_\delta(\theta^*) + \mathcal{T}_{\alpha-\delta}(\theta^*). && (5.23)\end{aligned}$$

The theorem finally follows by invoking the nondegeneracy condition, which ensures $\mathbb{E}[g_{\theta^*}(X_1,Y_1)] = 0$, so we have shown $0 \in \mathcal{R}_\delta(\theta^*) + \mathcal{T}_{\alpha-\delta}(\theta^*)$.

5.6.2 Asymptotic counterpart of Theorem 5.2.1

The following is an asymptotic counterpart of Theorem 5.2.1 that uses the central limit theorem in the confidence set construction. We note the error budget splitting used in Theorem 5.2.1 is in fact not necessary, but we believe that it facilitates exposition when presenting nonasymptotic guarantees. The asymptotic result below is stated without the splitting of the error budget.

Theorem 5.6.1 (Convex estimation: asymptotic version)**.** *Suppose that the convex estimation problem is nondegenerate as in* (5.4) *and that* $\frac{n}{N} \to p$*, for some* $p \in (0,1)$*. Fix* $\alpha \in (0,1)$*. For all* $\theta \in \mathbb{R}^p$*, define*

$$\hat{\boldsymbol{\Delta}}^f(\theta) = \frac{1}{n}\sum_{i=1}^n \left(g_\theta(X_i,Y_i) - g_\theta(X_i,f_i)\right); \quad \hat{g}^f(\theta) = \frac{1}{N}\sum_{i=1}^N g_\theta(\widetilde{X}_i,\tilde{f}_i).$$

Further, denoting by $g_{\theta,j}(x,y)$ *the* j*-th coordinate of* $g_\theta(x,y)$*, let*

$$\hat{\sigma}^2_{\Delta,j}(\theta) = \frac{1}{n}\sum_{i=1}^n \left(g_{\theta,j}(X_i,Y_i) - g_{\theta,j}(X_i,f_i) - \hat{\boldsymbol{\Delta}}^f_j(\theta)\right)^2; \quad \hat{\sigma}^2_{g,j}(\theta) = \frac{1}{N}\sum_{i=1}^N \left(g_{\theta,j}(\widetilde{X}_i,\tilde{f}_i) - \hat{g}^f_j(\theta)\right)^2,$$

for all $j \in [p]$*. Let* $w_{\alpha,j}(\theta) = z_{1-\alpha/(2p)}\sqrt{\frac{\hat{\sigma}^2_{\Delta,j}(\theta)}{n} + \frac{\hat{\sigma}^2_{g,j}(\theta)}{N}}$ *and*

$$\mathcal{C}_\alpha^{\mathrm{PP}} = \left\{\theta : |\hat{\boldsymbol{\Delta}}^f_j(\theta) + \hat{g}^f_j(\theta)| \leqslant w_{\alpha,j}(\theta),\ \forall j \in [p]\right\}.$$

Then,

$$\liminf_{n,N\to\infty} P(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}) \geqslant 1-\alpha. \tag{5.24}$$

Proof. We show that $\theta^* \notin \mathcal{C}_\alpha^{\mathrm{PP}}$ with probability at most α in the limit; that is,

$$\limsup_{n,N\to\infty} P\left(\left|\hat{\mathbf{\Delta}}_j^f(\theta^*) + \hat{g}_j^f(\theta^*)\right| > z_{1-\alpha/(2p)}\sqrt{\frac{\hat{\sigma}_{\Delta,j}^2(\theta^*)}{n} + \frac{\hat{\sigma}_{g,j}^2(\theta^*)}{N}},\ \forall j \in [p]\right) \leqslant \alpha.$$

For each $j \in [p]$, the central limit theorem implies that

$$\sqrt{n}(\hat{\mathbf{\Delta}}_j^f(\theta^*) - \mathbb{E}[\hat{\mathbf{\Delta}}_j^f(\theta^*)]) \Rightarrow \mathcal{N}(0, \sigma_{\Delta,j}^2(\theta^*)); \quad \sqrt{N}(\hat{g}_j^f(\theta^*) - \mathbb{E}[\hat{g}_j^f(\theta^*)]) \Rightarrow \mathcal{N}(0, \sigma_{g,j}^2(\theta^*)),$$

where $\sigma_{\Delta,j}^2(\theta^*)$ is the variance of $g_{\theta^*,j}(X_1, Y_1) - g_{\theta^*,j}(X_1, f_1)$ and $\sigma_{g,j}^2(\theta^*)$ is the variance of $g_{\theta^*,j}(X_1, f_1)$. Therefore, by Slutsky's theorem, we get

$$\begin{aligned}
&\sqrt{N}(\hat{\mathbf{\Delta}}_j^f(\theta^*) + \hat{g}_j^f(\theta^*) - \mathbb{E}[\hat{\mathbf{\Delta}}_j^f(\theta^*) + \hat{g}_j^f(\theta^*)]) \\
&= \sqrt{n}(\hat{\mathbf{\Delta}}_j^f(\theta^*) - \mathbb{E}[\hat{\mathbf{\Delta}}_j^f(\theta^*)])\sqrt{\frac{N}{n}} + \sqrt{N}(\hat{g}_j^f(\theta^*) - \mathbb{E}[\hat{g}_j^f(\theta^*)]) \\
&\Rightarrow \mathcal{N}\left(0, \frac{1}{p}\sigma_{\Delta,j}^2(\theta^*) + \sigma_{g,j}^2(\theta^*)\right).
\end{aligned}$$

This in turn implies

$$\limsup_{n,N\to\infty} P\left(\left|\hat{\mathbf{\Delta}}_j^f(\theta^*) + \hat{g}_j^f(\theta^*) - \mathbb{E}\left[\hat{\mathbf{\Delta}}_j^f(\theta^*) + \hat{g}_j^f(\theta^*)\right]\right| > z_{1-\alpha/(2p)}\frac{\hat{\sigma}_j}{\sqrt{N}}\right) \leqslant \frac{\alpha}{p}, \tag{5.25}$$

where $\hat{\sigma}_j^2$ is a consistent estimate of the variance $\frac{1}{p}\sigma_{\Delta,j}^2(\theta^*) + \sigma_{g,j}^2(\theta^*)$. We take $\hat{\sigma}_j^2 = \hat{\sigma}_{\Delta,j}^2(\theta^*)\frac{N}{n} + \hat{\sigma}_{g,j}^2(\theta^*)$; this estimate is consistent since the two terms are individually consistent estimates of the respective variances. Now notice that

$$\mathbb{E}\left[\hat{\mathbf{\Delta}}^f(\theta^*) + \hat{g}^f(\theta^*)\right] = \mathbb{E}\left[g_{\theta^*}(X_1, Y_1) - g_{\theta^*}(X_1, f_1) + g_{\theta^*}(\widetilde{X}_1, \tilde{f}_1)\right] = \mathbb{E}[g_{\theta^*}(X_1, Y_1)] = 0, \tag{5.26}$$

where the last step follows by the nondegeneracy condition. Putting together (5.25), (5.26), and the choice of $\hat{\sigma}_j$ derived above, and applying a union bound, we get

$$\begin{aligned}
&\limsup_{n,N\to\infty} P\left(\exists j \in [p] : \left|\hat{\mathbf{\Delta}}_j^f(\theta^*) + \hat{g}_j^f(\theta^*)\right| > z_{1-\alpha/(2p)}\sqrt{\frac{\hat{\sigma}_{\Delta,j}^2(\theta^*)}{n} + \frac{\hat{\sigma}_{g,j}^2(\theta^*)}{N}}\right) \\
&\leqslant \sum_{j=1}^p \limsup_{n,N\to\infty} P\left(\left|\hat{\mathbf{\Delta}}_j^f(\theta^*) + \hat{g}_j^f(\theta^*)\right| > z_{1-\alpha/(2p)}\sqrt{\frac{\hat{\sigma}_{\Delta,j}^2(\theta^*)}{n} + \frac{\hat{\sigma}_{g,j}^2(\theta^*)}{N}}\right) \\
&= \sum_{j=1}^p \limsup_{n,N\to\infty} P\left(\left|\hat{\mathbf{\Delta}}_j^f(\theta^*) + \hat{g}_j^f(\theta^*) - \mathbb{E}\left[\hat{\mathbf{\Delta}}_j^f(\theta^*) + \hat{g}_j^f(\theta^*)\right]\right| > z_{1-\alpha/(2p)}\hat{\sigma}_j\right) \\
&\leqslant \sum_{j=1}^p \frac{\alpha}{p} \\
&= \alpha.
\end{aligned}$$

□

5.6.3 Proof of Proposition 5.2.1

We show that the prediction-powered confidence set constructed in Algorithm 11 is a special case of the prediction-powered confidence set constructed in Theorem 5.6.1. The proof then follows directly by the guarantee of Theorem 5.6.1.

Since $g_\theta(y) = \theta - y$, we have

$$\hat{\boldsymbol{\Delta}}^f(\theta) \equiv \hat{\boldsymbol{\Delta}}^f = \frac{1}{n}\sum_{i=1}^n (f_i - Y_i); \quad \hat{g}^f(\theta) = \theta - \frac{1}{N}\sum_{i=1}^N \tilde{f}_i.$$

Therefore, the set $\mathcal{C}_\alpha^{\mathrm{PP}}$ from Theorem 5.6.1 can be written as

$$\mathcal{C}_\alpha^{\mathrm{PP}} = \left\{\theta : \left|\theta - \frac{1}{N}\sum_{i=1}^N \tilde{f}_i + \frac{1}{n}\sum_{i=1}^n (f_i - Y_i)\right| \leqslant w_\alpha(\theta)\right\} = \left(\frac{1}{N}\sum_{i=1}^N \tilde{f}_i - \frac{1}{n}\sum_{i=1}^n (f_i - Y_i) \pm w_\alpha(\theta)\right).$$

This is exactly the set constructed in Algorithm 11, which completes the proof.

5.6.4 Proof of Proposition 5.2.2

Like in the proof of Proposition 5.2.1, we proceed by showing that the prediction-powered confidence set constructed in Algorithm 12 is a special case of the prediction-powered confidence set constructed in Theorem 5.6.1. Then, we simply invoke Theorem 5.6.1.

Since $g_\theta(y) = -q + \mathbf{1}\{y \leqslant \theta\}$, we have

$$\hat{\boldsymbol{\Delta}}^f(\theta) = \frac{1}{n}\sum_{i=1}^n \left(\mathbf{1}\{Y_i \leqslant \theta\} - \mathbf{1}\{f_i \leqslant \theta\}\right); \quad \hat{g}^f(\theta) = -q + \hat{F}(\theta),$$

where $\hat{F}(\theta) = \frac{1}{N}\sum_{i=1}^N \mathbf{1}\left\{\tilde{f}_i \leqslant \theta\right\}$. Therefore, the set $\mathcal{C}_\alpha^{\mathrm{PP}}$ from Theorem 5.6.1 can be written as

$$\begin{aligned}\mathcal{C}_\alpha^{\mathrm{PP}} &= \left\{\theta : \left|\frac{1}{n}\sum_{i=1}^n \left(\mathbf{1}\{Y_i \leqslant \theta\} - \mathbf{1}\{f_i \leqslant \theta\}\right) - q + \hat{F}(\theta)\right| \leqslant w_\alpha(\theta)\right\}\\ &= \left\{\theta : \left|\hat{F}(\theta) + \hat{\boldsymbol{\Delta}}^f(\theta) - q\right| \leqslant w_\alpha(\theta)\right\}.\end{aligned}$$

This is exactly the set constructed in Algorithm 12. Therefore, the guarantee of Proposition 5.2.2 follows by the guarantee of Theorem 5.6.1.

5.6.5 Proof of Proposition 5.2.3

The proof follows a similar pattern as the previous two propositions, by arguing that the prediction-powered confidence set constructed in Algorithm 13 is a special case of the prediction-powered confidence set constructed in Theorem 5.6.1.

Since $g_\theta(x, y) = x(\mu_\theta(x) - y)$, we have

$$\hat{\mathbf{\Delta}}^f(\theta) \equiv \hat{\mathbf{\Delta}}^f = \frac{1}{n}\sum_{i=1}^{n} X_i(f_i - Y_i); \quad \hat{g}^f(\theta) = \frac{1}{N}\sum_{i=1}^{N} \widetilde{X}_i(\mu_\theta(\widetilde{X}_i) - \tilde{f}_i).$$

These quantities are explicitly computed in Algorithm 13. Moreover, the set $\mathcal{C}_\alpha^{\mathrm{PP}}$ constructed in Algorithm 13 exactly follows the recipe of Theorem 5.6.1, so the proof immediately follows.

5.6.6 Proof of Proposition 5.2.4

For linear regression, we can derive more powerful prediction-powered confidence intervals than those implied by Theorem 5.2.1 by exploiting the linearity of the least-squares estimator.

Recall that Theorem 5.6.1 assumes that $\frac{n}{N} \to p$, for some fraction $p \in (0, 1)$.

Theorem 3 of White [180] implies that

$$\sqrt{n}(\hat{\mathbf{\Delta}}^f - \mathbf{\Delta}^f) \Rightarrow \mathcal{N}(0, W); \quad \sqrt{N}(\tilde{\theta}^f - \theta^f) \Rightarrow \mathcal{N}(0, W'),$$

for appropriately defined coviariance matrices W and W', where $\theta^f = (\mathbb{E}[X_1 X_1^\top])^{-1}\,\mathbb{E}[X_1 f_1]$ and $\mathbf{\Delta}^f = (\mathbb{E}[X_1 X_1^\top])^{-1}\,\mathbb{E}[X_1(f_1 - Y_1)]$. With this, we can write the target estimand as $\theta^* = (\mathbb{E}[X_1 X_1^\top])^{-1}\,\mathbb{E}[X_1 Y_1] = \theta^f - \mathbf{\Delta}^f$.

Combining Theorem 3 of White with Slutsky's theorem, we get

$$\sqrt{N}(\hat{\theta}^{\mathrm{PP}} - \theta^*) = \sqrt{N}(\tilde{\theta}^f - \theta^f) - \sqrt{n}(\hat{\mathbf{\Delta}}^f - \mathbf{\Delta}^f)\sqrt{\frac{N}{n}} \Rightarrow \mathcal{N}\left(0, W\frac{1}{p} + W'\right).$$

White also shows that V and $\tilde{V}$, as defined in Algorithm 14, are consistent estimates of W and W', respectively. Therefore, $\hat{\theta}^{\mathrm{PP}}$ is asymptotically normal and consistent, and we have a consistent estimate of its covariance. In particular,

$$V_{j^*j^*}\frac{N}{n} + \tilde{V}_{j^*j^*} \to W_{j^*j^*}\frac{1}{p} + W'_{j^*j^*}.$$

This means that we can construct asymptotically valid confidence intervals via a normal approximation by choosing width $z_{1-\alpha/2}\sqrt{V_{j^*j^*}\frac{N}{n} + \tilde{V}_{j^*j^*}}\sqrt{\frac{1}{N}} = z_{1-\alpha/2}\sqrt{\frac{V_{j^*j^*}}{n} + \frac{\tilde{V}_{j^*j^*}}{N}}$, and this is precisely what Algorithm 14 accomplishes.

5.6.7 Proof of Theorem 5.4.1

Define

$$L(\theta) = \mathbb{E}[\ell_\theta(X_1, Y_1)], \quad L^f(\theta) = \mathbb{E}[\ell_\theta(X_1, f_1)].$$

By the definition of θ^*, we have

$$\begin{aligned}\tilde{L}^f(\theta^*) &= (\tilde{L}^f(\theta^*) - L(\theta^*)) + (L(\theta^*) - L(\tilde{\theta}^f)) + (L(\tilde{\theta}^f) - \tilde{L}^f(\tilde{\theta}^f)) + \tilde{L}^f(\tilde{\theta}^f) \\ &\leqslant (\tilde{L}^f(\theta^*) - L(\theta^*)) + (L(\tilde{\theta}^f) - \tilde{L}^f(\tilde{\theta}^f)) + \tilde{L}^f(\tilde{\theta}^f).\end{aligned}$$

By applying the validity of the confidence bounds, a union bound implies that with probability $1-\alpha$ we have

$$\begin{aligned}\tilde{L}^f(\theta^*) &\leqslant (L^f(\theta^*) - L(\theta^*)) + (L(\tilde{\theta}^f) - L^f(\tilde{\theta}^f)) + \tilde{L}^f(\tilde{\theta}^f) + \mathcal{T}^u_{\frac{\alpha-\delta}{2}}(\theta^*) - \mathcal{T}^l_{\frac{\alpha-\delta}{2}}(\tilde{\theta}^f) \\ &= -\boldsymbol{\Delta}^f(\theta^*) + \boldsymbol{\Delta}^f(\tilde{\theta}^f) + \tilde{L}^f(\tilde{\theta}^f) + \mathcal{T}^u_{\frac{\alpha-\delta}{2}}(\theta^*) - \mathcal{T}^l_{\frac{\alpha-\delta}{2}}(\tilde{\theta}^f) \\ &\leqslant -\mathcal{R}^l_{\delta/2}(\theta^*) + \mathcal{R}^u_{\delta/2}(\tilde{\theta}^f) + \tilde{L}^f(\tilde{\theta}^f) + \mathcal{T}^u_{\frac{\alpha-\delta}{2}}(\theta^*) - \mathcal{T}^l_{\frac{\alpha-\delta}{2}}(\tilde{\theta}^f).\end{aligned}$$

Therefore, with probability $1-\alpha$ we have that $\theta^* \in \mathcal{C}^{\mathrm{PP}}_\alpha$, as desired.

5.6.8 Proof of Theorem 5.4.2

Notice that we can write $\mathbb{E}_{\mathbb{Q}_Y}[\nu(Y)] = \nu^\top \mathbb{Q}_Y$, where on the right-hand side we are treating $\nu = (\nu(1), \ldots, \nu(K))$ and $\mathbb{Q}_Y = (\mathbb{Q}_Y(1), \ldots, \mathbb{Q}_Y(K))$ as vectors of length K. We can write similar expressions for $\mathbb{Q}_f, \widehat{\mathbb{Q}}_Y$, etc. Using this notation, by triangle inequality we have

$$|\theta^* - \nu^\top \widehat{\mathbb{Q}}_Y| = |\nu^\top \mathbb{Q}_Y - \nu^\top \widehat{\mathbb{Q}}_Y| \leqslant \left|\nu^\top \widehat{\mathcal{K}}^{-1}(\mathbb{Q}_f - \widehat{\mathbb{Q}}_f)\right| + \left|\nu^\top \mathcal{K}^{-1}\mathbb{Q}_f - \nu^\top \widehat{\mathcal{K}}^{-1}\mathbb{Q}_f\right|. \tag{5.27}$$

We bound the first term using Hölder's inequality,

$$\left|\nu^\top \widehat{\mathcal{K}}^{-1}(\mathbb{Q}_f - \widehat{\mathbb{Q}}_f)\right| \leqslant \|\nu^\top \widehat{\mathcal{K}}^{-1}\|_1 \|\mathbb{Q}_f - \widehat{\mathbb{Q}}_f\|_\infty. \tag{5.28}$$

For the second term, we write

$$\left|\nu^\top \mathcal{K}^{-1}\mathbb{Q}_f - \nu^\top \widehat{\mathcal{K}}^{-1}\mathbb{Q}_f\right| = \left|\nu^\top \widehat{\mathcal{K}}^{-1}(\widehat{\mathcal{K}} - \mathcal{K})\mathcal{K}^{-1}\mathbb{Q}_f\right|. \tag{5.29}$$

In the above equation, the factor on the right, $\mathcal{K}^{-1}\mathbb{Q}_f$, is exactly equal to $\mathbb{Q}_Y$, and thus lives on the simplex, which we denote by Δ. Using this fact and Hölder's inequality,

$$\left|\nu^\top \widehat{\mathcal{K}}^{-1}(\widehat{\mathcal{K}} - \mathcal{K})\mathcal{K}^{-1}\mathbb{Q}_f\right| \leqslant \sup_{q \in \Delta} \left|\nu^\top \widehat{\mathcal{K}}^{-1}(\widehat{\mathcal{K}} - \mathcal{K})q\right| \leqslant \left\|\nu^\top \widehat{\mathcal{K}}^{-1}\right\|_1 \sup_{q \in \Delta} \left\|(\widehat{\mathcal{K}} - \mathcal{K})q\right\|_\infty. \tag{5.30}$$

Next, we have

$$\sup_{q\in\Delta} \|(\widehat{\mathcal{K}} - \mathcal{K})q\|_\infty = \max_{k\in[K]} \|\widehat{\mathcal{K}}_k - \mathcal{K}_k\|_\infty, \tag{5.31}$$

where $\mathcal{K}_k$ indexes the k-th column of $\mathcal{K}$. This yields the expression

$$\left\|\nu^\top \widehat{\mathcal{K}}^{-1}\right\|_1 \sup_{q\in\Delta} \left\|(\widehat{\mathcal{K}} - \mathcal{K})q\right\|_\infty = \left\|\nu^\top \widehat{\mathcal{K}}^{-1}\right\|_1 \max_{k\in[K]} \|\widehat{\mathcal{K}}_k - \mathcal{K}_k\|_\infty. \tag{5.32}$$

Putting everything together and going back to (5.27), we have

$$|\nu^\top \mathbb{Q}_Y - \nu^\top \widehat{\mathbb{Q}}_Y| \leqslant \|\nu^\top \widehat{\mathcal{K}}^{-1}\|_1 \left(\|\mathbb{Q}_f - \widehat{\mathbb{Q}}_f\|_\infty + \max_{k\in[K]} \|\widehat{\mathcal{K}}_k - \mathcal{K}_k\|_\infty\right). \tag{5.33}$$

Since $\|\nu^\top \widehat{\mathcal{K}}^{-1}\|_1$ can be evaluated empirically, it remains to bound the distributional distances $\|\mathbb{Q}_f - \widehat{\mathbb{Q}}_f\|_\infty$ and $\max_{k\in[K]} \|\widehat{\mathcal{K}}_k - \mathcal{K}_k\|_\infty$.

For the first term, we can simply apply the DKWM inequality [52, 119], which gives

$$\|\mathbb{Q}_f - \widehat{\mathbb{Q}}_f\|_\infty \leqslant \sqrt{\frac{2}{N} \log \frac{2}{\alpha - \delta}} \tag{5.34}$$

with probability $1 - (\alpha - \delta)$. See [27] for details.

For the second term, $\max_{k\in[K]} \|\widehat{\mathcal{K}}_k - \mathcal{K}_k\|_\infty$, since we only have n samples for estimation, we use a more adaptive concentration result. In particular, for each $l, k \in [K]$, $n(k)\widehat{\mathcal{K}}_{l,k}$ (conditional on the k-th column) follows a binomial distribution with $n(k)$ samples and success probability $\mathcal{K}_{l,k}$. Therefore, if we let

$$C_{l,k} = \left\{p : n(k)\widehat{\mathcal{K}}_{l,k} \in \left(F^{-1}_{\mathrm{Binom}(n(k),p)}\left(\frac{\delta}{2K^2}\right), F^{-1}_{\mathrm{Binom}(n(k),p)}\left(1 - \frac{\delta}{2K^2}\right)\right)\right\},$$

where $F_{\mathrm{Binom}(n(k),p)}$ denotes the Binomial CDF, then by a union bound:

$$P\left(\max_{k\in[K]} \|\widehat{\mathcal{K}}_k - \mathcal{K}_k\|_\infty \geqslant \max_{l,k\in[K]} \max_{p\in C_{l,k}} |\widehat{\mathcal{K}}_{l,k} - p|\right) \leqslant \delta. \tag{5.35}$$

Combining equations (5.33), (5.34) and (5.35) yields the final result.

Bibliography

[1] Alekh Agarwal and Ofer Dekel. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Conference on Learning Theory*, pages 28–40, 2010.

[2] Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. Inference on winners. Technical report, National Bureau of Economic Research, 2019.

[3] Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *arXiv preprint arXiv:2301.09633*, 2023.

[4] David Azriel, Lawrence D. Brown, Michael Sklar, Richard Berk, Andreas Buja, and Linda Zhao. Semi-supervised linear regression. *Journal of the American Statistical Association*, 117(540):2238–2251, 2022.

[5] Peter L. Bartlett. Learning with a slowly changing distribution. In *Proceedings of the Fifth Annual ACM Conference on Computational Learning Theory (COLT)*, pages 243–252, 1992.

[6] Peter L. Bartlett, Shai Ben-David, and Sanjeev R. Kulkarni. Learning changing concepts by exploiting the structure of change. *Machine Learning*, 41(2):153–174, 2000.

[7] Raef Bassily and Yoav Freund. Typical stability. *arXiv preprint arXiv:1604.03336*, 2016.

[8] Yahav Bechavod, Katrina Ligett, Steven Wu, and Juba Ziani. Gaming helps! learning from strategic interactions in natural dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 1234–1242, 2021.

[9] Yoav Benjamini, Yotam Hechtlinger, and Philip B Stark. Confidence intervals for selected parameters. *arXiv preprint arXiv:1906.00505*, 2019.

[10] George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

[11] Vidmantas Bentkus. On hoeffding's inequalities. *Annals of Probability*, 32(2):1650–1673, 2004.

[12] Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, pages 802–837, 2013.

[13] Sergei Bernstein. On a modification of Chebyshev's inequality and of the error formula of Laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49, 1924.

[14] Nan Bi, Jelena Markovic, Lucy Xia, and Jonathan Taylor. Inferactive data analysis. *Scandinavian Journal of Statistics*, 47(1):212–249, 2020.

[15] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[16] Isabell Bludau, Sander Willems, Wen-Feng Zeng, Maximilian T Strauss, Fynn M Hansen, Maria C Tanzer, Ozge Karayel, Brenda A Schulman, and Matthias Mann. The structural context of posttranslational modifications at a proteome-wide scale. *PLoS Biology*, 20(5):e3001636, 2022.

[17] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

[18] F Jay Breidt and Jean D Opsomer. Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2):190–205, 2017.

[19] Frank Bretz, Alan Genz, and Ludwig A. Hothorn. On the numerical availability of multiple comparison procedures. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 43(5):645–656, 2001.

[20] Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *International Conference on Artificial Intelligence and Statistics*, pages 6045–6061. PMLR, 2022.

[21] Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. *Journal of Machine Learning Research*, 13(Sep):2617–2654, 2012.

[22] Andreas Buja, Lawrence Brown, Richard Berk, Edward George, Emil Pitkin, Mikhail Traskin, Kai Zhang, and Linda Zhao. Models as approximations I. *Statistical Science*, 34(4):523–544, 2019.

[23] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

[24] Jenna Burrell, Zoe Kahn, Anne Jonas, and Daniel Griffin. When users control the algorithms: values expressed in practices on Twitter. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–20, 2019.

[25] Jenna Burrell, Zoe Kahn, Anne Jonas, and Daniel Griffin. When users control the algorithms: Values expressed in practices on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, 3:19, 2019.

[26] Adriana Camacho and Emily Conover. Manipulation of social program eligibility. *American Economic Journal: Economic Policy*, 3(2):41–65, 2011.

[27] Clément L Canonne. A short note on learning discrete distributions. *arXiv preprint arXiv:2002.11457*, 2020.

[28] Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective.* Chapman and Hall/CRC, 2006.

[29] Claes M Cassel, Carl E Särndal, and Jan H Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620, 1976.

[30] Abhishek Chakrabortty and Tianxi Cai. Efficient and adaptive linear regression in semi-supervised settings. *Annals of Statistics*, 46(4):1541–1572, 2018.

[31] Abhishek Chakrabortty, Guorong Dai, and Raymond J Carroll. Semi-supervised quantile estimation: Robust and efficient inference in high dimensional settings. *arXiv preprint arXiv:2201.10208*, 2022.

[32] Alan Chan. Scoring rules for performative binary prediction. *arXiv preprint arXiv:2207.02847*, 2022.

[33] Jinbo Chen and Norman E Breslow. Semiparametric efficient estimation for the auxiliary outcome problem with the conditional mean model. *Canadian Journal of Statistics*, 32(4):359–372, 2004.

[34] Julie Yujie Chen. Thrown under the bus and outrunning it! The logic of Didi and taxi drivers' labour and activism in the on-demand economy. *New Media & Society*, 20(8):2691–2711, 2018.

[35] Song Xi Chen, Denis H Y Leung, and Jing Qin. Information recovery in a study with surrogate endpoints. *Journal of the American Statistical Association*, 98(464):1052–1062, 2003.

[36] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[37] Xiaohong Chen, Han Hong, and Elie Tamer. Measurement error models with auxiliary data. *The Review of Economic Studies*, 72(2):343–366, 2005.

[38] Yiling Chen, Yang Liu, and Chara Podimata. Learning strategy-aware linear classifiers. *Advances in Neural Information Processing Systems*, 33:15265–15276, 2020.

[39] Kenneth L Clarkson. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. *ACM Transactions on Algorithms*, 6(4):1–30, 2010.

[40] Kelley Cotter. Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram. *New Media & Society*, 21(4):895–913, 2019.

[41] Elliot Creager and Richard Zemel. Online algorithmic recourse by collective action. *ICML Workshop on Algorithmic Recourse*, 2021.

[42] Rachel Cummings, Katrina Ligett, Kobbi Nissim, Aaron Roth, and Zhiwei Steven Wu. Adaptive learning with robust generalization guarantees. In *Conference on Learning Theory (COLT)*, pages 772–814, 2016.

[43] Joshua Cutler, Mateo Díaz, and Dmitriy Drusvyatskiy. Stochastic approximation with decision-dependent distributions: asymptotic normality and optimality. *arXiv preprint arXiv:2207.04173*, 2022.

[44] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 99–108, 2004.

[45] Sarah Dean, Mihaela Curmei, Lillian J Ratliff, Jamie Morgenstern, and Maryam Fazel. Multi-learner risk reduction under endogenous participation dynamics. *arXiv preprint arXiv:2206.02667*, 2022.

[46] Drew Desilver. On Election Day, most voters use electronic or optical-scan ballots. *Fact Tank*, 2016.

[47] Thorsten Dickhaus. *Simultaneous statistical inference: with applications in the life sciences*. Springer, 2014.

[48] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34:6478–6490, 2021.

[49] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70. ACM, 2018.

[50] Roy Dong, Heling Zhang, and Lillian Ratliff. Approximate regions of attraction in learning with decision-dependent distributions. In *International Conference on Artificial Intelligence and Statistics*, pages 11172–11184. PMLR, 2023.

[51] Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 2022.

[52] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

[53] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2350–2358, 2015.

[54] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

[55] Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*, volume 9. Now Publishers, Inc., 2014.

[56] Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *IEEE 51st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 51–60, 2010.

[57] Bradley Efron. *Exponential Families in Theory and Practice.* Cambridge University Press, 2022.

[58] Bradley Efron. *Exponential Families in Theory and Practice.* Cambridge University Press, 2022.

[59] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *COLT*, volume 2, pages 255–270. Springer, 2002.

[60] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[61] William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

[62] 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*, pages 385–394, 2005.

[63] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

[64] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys (CSUR)*, 46(4):1–37, 2014.

[65] Alan Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1(2):141–149, 1992.

[66] Alan Genz and Frank Bretz. Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts. *Journal of Statistical Computation and Simulation*, 63(4):103–117, 1999.

[67] Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe feature elimination for the LASSO and sparse supervised learning problems. *arXiv preprint arXiv:1009.4219*, 2010.

[68] Jelle Goeman and Aldo Solari. Conditional versus unconditional approaches to selective inference. *arXiv preprint arXiv:2207.13480*, 2022.

[69] Paula Gradu, Tijana Zrnic, Yixin Wang, and Michael I Jordan. Valid inference after causal discovery. *arXiv preprint arXiv:2208.05949*, 2022.

[70] Mary L Gray and Siddharth Suri. *Ghost work: How to stop Silicon Valley from building a new global underclass.* Eamon Dolan Books, 2019.

[71] Kathrin Grosse, Taesung Lee, Battista Biggio, Youngja Park, Michael Backes, and Ian Molloy. Backdoor smoothing: Demystifying backdoor attacks on deep neural networks. *Computers & Security*, 120:102814, 2022.

[72] Wei Guo, Benedetta Tondi, and Mauro Barni. An overview of backdoor attacks against deep neural networks and possible defences. *IEEE Open Journal of Signal Processing*, 3:261–287, 2022.

[73] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(Mar):1157–1182, 2003.

[74] Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner. Performative power. *arXiv preprint arXiv:2203.17232*, 2022.

[75] Moritz Hardt, Eric Mazumdar, Celestine Mendler-Dünner, and Tijana Zrnic. Algorithmic collective action in machine learning. *International Conference on Machine Learning*, 2023.

[76] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the ACM Conference on Innovations in Theoretical Computer Science*, pages 111–122, 2016.

[77] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[78] Kieran Healy. The performativity of networks. *European Journal of Sociology/Archives Européennes de Sociologie*, 56(2):175–205, 2015.

[79] Yosef Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

[80] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[81] Gerhard Hommel. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75(2):383–386, 1988.

[82] Torsten Hothorn, Frank Bretz, and Peter Westfall. Simultaneous inference in general parametric models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(3):346–363, 2008.

[83] Jue Hou, Zijian Guo, and Tianxi Cai. Surrogate assisted semi-supervised inference for high dimensional risk prediction. *arXiv preprint arXiv:2105.01264*, 2021.

[84] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency*, pages 259–268, 2019.

[85] Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical sciences*. Cambridge University Press, 2015.

[86] 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*, pages 4641–4650. PMLR, 2021.

[87] 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*, pages 3998–4035. PMLR, 2022.

[88] Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. Regret minimization with performative feedback. In *International Conference on Machine Learning*, pages 9760–9785, 2022.

[89] Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, pages 427–435, 2013.

[90] Kameni Florentin Flambeau Jiechieu and Norbert Tsopze. Skills prediction based on multi-label resume classification using cnn with model predictions explanation. *Neural Computing and Applications*, 33:5069–5087, 2021.

[91] Kun Jin, Tongxin Yin, Zhongzhu Chen, Zeyu Sun, Xueru Zhang, Yang Liu, and Mingyan Liu. Performative federated learning. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*, 2023.

[92] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[93] Kaggle. Give me some credit. `https://www.kaggle.com/c/GiveMeSomeCredit/data`, 2012.

[94] Nathan Kallus and Xiaojie Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *arXiv preprint arXiv:2003.12408*, 2020.

[95] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811, 2016.

[96] Michael P Kim and Juan C Perdomo. Making decisions under outcome performativity. *arXiv preprint arXiv:2210.01745*, 2022.

[97] Danijel Kivaranovic and Hannes Leeb. A (tight) upper bound for the length of confidence intervals with conditional coverage. *arXiv preprint arXiv:2007.12448*, 2020.

[98] Danijel Kivaranovic and Hannes Leeb. On the length of post-model-selection confidence intervals conditional on polyhedral constraints. *Journal of the American Statistical Association*, 116(534):845–857, 2021.

[99] Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the ACM Conference on Economics and Computation (EC)*, pages 825–844, 2019.

[100] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690, 2008.

[101] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50, 1978.

[102] Arun K Kuchibhotla, Lawrence D Brown, Andreas Buja, and Junhui Cai. All of linear regression. *arXiv preprint arXiv:1910.06386*, 2019.

[103] Anthony Kuh, Thomas Petsche, and Ronald L Rivest. Learning time-varying concepts. In *Advances in Neural Information Processing Systems (NIPS)*, pages 183–189, 1991.

[104] Bogdan Kulynych. Causal prediction can induce performative stability. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022.

[105] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms.* Cambridge University Press, 2020.

[106] Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[107] Jason D Lee and Jonathan E Taylor. Exact post model selection inference for marginal screening. In *Advances in Neural Information Processing Systems (NIPS)*, pages 136–144, 2014.

[108] Sagi Levanon and Nir Rosenfeld. Strategic classification made practical. In *Proceedings of the 38th International Conference on Machine Learning*, pages 6243–6253, 2021.

[109] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[110] Qiang Li and Hoi-To Wai. State dependent performative prediction with stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3164–3186. PMLR, 2022.

[111] Qiang Li, Chung-Yiu Yau, and Hoi-To Wai. Multi-agent performative prediction with greedy deployment and consensus seeking agents. *arXiv preprint arXiv:2209.03811*, 2022.

[112] Mark Lindeman and Philip B Stark. A gentle introduction to risk-limiting audits. *IEEE Security & Privacy*, 10(5):42–49, 2012.

[113] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, pages 3122–3130. PMLR, 2018.

[114] Roderick JA Little and Donald B Rubin. *Statistical Analysis with Missing Data*, volume 793. John Wiley & Sons, 2019.

[115] Donald A MacKenzie, Fabian Muniesa, and Lucia Siu. *Do Economists Make Markets?: On the Performativity of Economics.* Princeton University Press, 2007.

[116] Chinmay Maheshwari, Chih-Yuan Chiu, Eric Mazumdar, Shankar Sastry, and Lillian Ratliff. Zeroth-order methods for convex-concave min-max problems: Applications to decision-dependent risk minimization. In *International Conference on Artificial Intelligence and Statistics*, pages 6702–6734. PMLR, 2022.

[117] Debmalya Mandal, Stelios Triantafyllou, and Goran Radanovic. Performative reinforcement learning. *arXiv preprint arXiv:2207.00046*, 2022.

[118] Jelena Markovic and Jonathan Taylor. Bootstrap inference after using multiple queries for model selection. *arXiv preprint arXiv:1612.07811*, 2016.

[119] Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, pages 1269–1283, 1990.

[120] Nikolai Matni and Stephen Tu. A tutorial on concentration bounds for system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3741–3749. IEEE, 2019.

[121] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proc.* 20*th International Conference on Artificial Intelligence and Statistics*, 2017.

[122] Celestine Mendler-Dünner, Frances Ding, and Yixin Wang. Anticipating performativity by predicting from predictions. *Advances in Neural Information Processing Systems*, 35:31171–31185, 2022.

[123] Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33:4929–4939, 2020.

[124] John Miller, Chloe Hsu, Jordan Troutman, Juan Perdomo, Tijana Zrnic, Lydia Liu, Yu Sun, Ludwig Schmidt, and Moritz Hardt. Whynot, 2020.

[125] John P Miller, Juan C Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, pages 7710–7720, 2021.

[126] Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency*, pages 230–239, 2019.

[127] Mehrnaz Mofakhami, Ioannis Mitliagkas, and Gauthier Gidel. Performative prediction with neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 11079–11093. PMLR, 2023.

[128] Marieke Möhlmann and Lior Zalmanson. Hands on the wheel: Navigating algorithmic management and uber drivers'. In *Autonomy', in proceedings of the international conference on information systems (ICIS), Seoul South Korea*, pages 10–13, 2017.

[129] Adhyyan Narang, Evan Faulkner, Dmitriy Drusvyatskiy, Maryam Fazel, and Lillian Ratliff. Learning in stochastic monotone games with decision-dependent data. In *International Conference on Artificial Intelligence and Statistics*, pages 5891–5912. PMLR, 2022.

[130] Mancur Olson. *The logic of collective action: public goods and the theory of groups.* Number 124 in Harvard economic studies. Harvard Univ. Press, 1965.

[131] Robert J Olson, Alexi Shalapyonok, and Heidi M Sosik. An automated submersible flow cytometer for analyzing pico-and nanophytoplankton: FlowCytobot. *Deep Sea Research Part I: Oceanographic Research Papers*, 50(2):301–315, 2003.

[132] Eric C Orenstein, Oscar Beijbom, Emily E Peacock, and Heidi M Sosik. WHOI-plankton-a large scale fine grained visual recognition benchmark dataset for plankton classification. *arXiv preprint arXiv:1510.00745*, 2015.

[133] Snigdha Panigrahi, Jelena Markovic, and Jonathan Taylor. An MCMC-free approach to post-selective inference. *arXiv preprint arXiv:1703.06154*, 2017.

[134] Snigdha Panigrahi and Jonathan Taylor. Approximate selective inference via maximum likelihood. *arXiv preprint arXiv:1902.07884*, 2019.

[135] Judea Pearl. *Causality.* Cambridge University Press, 2009.

[136] Margaret Sullivan Pepe. Inference using surrogate outcome data and a validation sample. *Biometrika*, 79(2):355–365, 1992.

[137] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609, 2020.

[138] Georgios Piliouras and Fang-Yi Yu. Multi-agent performative prediction: From global stability and optimality to chaos. *arXiv preprint arXiv:2201.10483*, 2022.

[139] José Pombal, Pedro Saleiro, Mário AT Figueiredo, and Pedro Bizarro. Prisoners of their own devices: How models induce data bias in performative prediction. *arXiv preprint arXiv:2206.13183*, 2022.

[140] Hatim A. Rahman. The invisible cage: Workers' reactivity to opaque algorithmic evaluations. *Administrative Science Quarterly*, 66(4):945–988, 2021.

[141] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1571–1578, 2012.

[142] Daniel G Rasines and G Alastair Young. Splitting strategies for post-selection inference. *arXiv preprint arXiv:2102.02159*, 2021.

[143] Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. Weatherbench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020.

[144] Mitas Ray, Lillian J Ratliff, Dmitriy Drusvyatskiy, and Maryam Fazel. Decision-dependent risk minimization in geometrically decaying dynamic environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8081–8088, 2022.

[145] James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

[146] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

[147] Tim Roughgarden. Algorithmic game theory. *Communications of the ACM*, 53(7):78–86, 2010.

[148] Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model Assisted Survey Sampling*. Springer Science & Business Media, 1992.

[149] Henry Scheffe. *The Analysis of Variance*. John Wiley & Sons, 1999.

[150] Juliet Schor. *After the gig: How the sharing economy got hijacked and how to win it back*. University of California Press, 2021.

[151] Moshe Shaked and J George Shanthikumar. *Stochastic orders*. Springer, 2007.

[152] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

[153] Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory*, pages 3–24, 2013.

[154] Yonadav Shavit, Benjamin Edelman, and Brian Axelrod. Learning from strategic agents: Accuracy, improvement, and causality. *arXiv preprint arXiv:2002.10066*, 2020.

[155] Meital Ben Sinai, Nimrod Partush, Shir Yadid, and Eran Yahav. Exploiting social navigation. *arXiv preprint arXiv:1410.0151*, 2014.

[156] Marilyn Strathern. Improving ratings: Audit in the british university system. *European Review*, 5(3):305–321, 1997.

[157] Kunal Talwar, Abhradeep Guha Thakurta, and Li Zhang. Nearly optimal private LASSO. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3025–3033, 2015.

[158] Xiaoying Tian, Nan Bi, and Jonathan Taylor. MAGIC: a general, powerful and tractable method for selective inference. *arXiv preprint arXiv:1607.02630*, 2016.

[159] Xiaoying Tian and Jonathan Taylor. Selective inference with a randomized response. *Annals of Statistics*, 46(2):679–710, 2018.

[160] Zhiyi Tian, Lei Cui, Jie Liang, and Shui Yu. A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Computing Surveys*, 55(8), 2022.

[161] Xiaoying Tian Harris, Snigdha Panigrahi, Jelena Markovic, Nan Bi, and Jonathan Taylor. Selective sampling after solving a convex problem. *arXiv preprint arXiv:1609.05609*, 2016.

[162] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[163] Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):245–266, 2012.

[164] Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.

[165] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, 2021.

[166] Eeshit Dhaval Vaishnav, Carl G de Boer, Jennifer Molinet, Moran Yassour, Lin Fan, Xian Adiconis, Dawn A Thompson, Joshua Z Levin, Francisco A Cubillos, and Aviv Regev. The evolution, evolvability and engineering of gene regulatory DNA. *Nature*, 603(7901):455–463, March 2022.

[167] Steven Vallas and Juliet B Schor. What do platforms do? Understanding the gig economy. *Annual Review of Sociology*, 46(1):273–294, 2020.

[168] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[169] Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2008.

[170] Nicholas Vincent and Brent Hecht. Can "conscious data contribution" help users to exert "data leverage" against technology companies? *Proc. ACM Hum.-Comput. Interact.*, 2021.

[171] Nicholas Vincent, Brent Hecht, and Shilad Sen. "Data strikes": evaluating the effectiveness of a new form of collective action against technology companies. In *The World Wide Web Conference*, 2019.

[172] Nicholas Vincent, Hanlin Li, Nicole Tilly, Stevie Chancellor, and Brent Hecht. Data leverage: A framework for empowering the public in its relationship with technology companies. In *Conference on Fairness, Accountability, and Transparency*, 2021.

[173] Heinrich von Stackelberg. *Market Structure and Equilibrium.* Springer Berlin Heidelberg, 2010.

[174] Indrė Žliobaitė. Learning under concept drift: An overview. *arXiv preprint arXiv:1010.4784*, 2010.

[175] Siruo Wang, Tyler H McCormick, and Jeffrey T Leek. Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117(48):30266–30275, 2020.

[176] Larry Wasserman and John Lafferty. Statistical analysis of semi-supervised regression. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[177] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *arXiv preprint arXiv:2010.09686*, 2020.

[178] Geoffrey I. Webb, Roy Hyde, Hong Cao, Hai Long Nguyen, and Francois Petitjean. Characterizing concept drift. *Data Mining and Knowledge Discovery*, 30(4):964–994, 2016.

[179] Sean Jeremy Westwood, Solomon Messing, and Yphtach Lelkes. Projecting confidence: How the probabilistic horse race confuses and demobilizes the public. *The Journal of Politics*, 82(4):1530–1544, 2020.

[180] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838, 1980.

[181] Alex J Wood, Mark Graham, Vili Lehdonvirta, and Isis Hjorth. Good gig, bad gig: autonomy and algorithmic control in the global gig economy. *Work, Employment and Society*, 33(1):56–75, 2019.

[182] Killian Wood, Gianluca Bianchin, and Emiliano Dall'Anese. Online projected gradient descent for stochastic optimization with decision-dependent distributions. *IEEE Control Systems Letters*, 6:1646–1651, 2021.

[183] Killian Wood and Emiliano Dall'Anese. Stochastic saddle point problems with decision-dependent distributions. *arXiv preprint arXiv:2201.02313*, 2022.

[184] Jamie Woodcock and Mark Graham. The gig economy. *A critical introduction. Cambridge: Polity*, 2019.

[185] Stephen J Wright and Benjamin Recht. *Optimization for data analysis.* Cambridge University Press, 2022.

[186] Changbao Wu and Randy R Sitter. A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453):185–193, 2001.

[187] Menggang Yu and Bin Nan. A revisit of semiparametric regression models with missing data. *Statistica Sinica*, pages 1193–1212, 2006.

[188] Yuqian Zhang and Jelena Bradic. High-dimensional semi-supervised learning: in search of optimal inference of the mean. *Biometrika*, 109(2):387–403, 2022.

[189] Yulai Zhao. Optimizing the performative risk under weak convexity assumptions. *arXiv preprint arXiv:2209.00771*, 2022.

[190] Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1):1–130, 2009.

[191] Xiaojin Jerry Zhu. Semi-supervised learning literature survey. 2005.

[192] Tijana Zrnic and William Fithian. Locally simultaneous inference. *arXiv preprint arXiv:2212.09009*, 2022.

[193] Tijana Zrnic and Michael I Jordan. Post-selection inference via algorithmic stability. *arXiv preprint arXiv:2011.09462*, 2020.

[194] Tijana Zrnic, Eric Mazumdar, Shankar Sastry, and Michael Jordan. Who leads and who follows in strategic classification? *Advances in Neural Information Processing Systems*, 34:15257–15269, 2021.