Plug-in Performative Optimization

Licong Lin [1] **Tijana Zrnic** [2]

Abstract

When predictions are performative, the choice of which predictor to deploy influences the distribution of future observations. The overarching goal in learning under performativity is to find a predictor that has low *performative risk*, that is, good performance on its induced distribution. One family of solutions for optimizing the performative risk, including bandits and other derivative-free methods, is agnostic to any structure in the performative feedback, leading to exceedingly slow convergence rates. A complementary family of solutions makes use of explicit *models* for the feedback, such as best-response models in strategic classification, enabling faster rates. However, these rates critically rely on the feedback model being correct. In this work we study a general protocol for making use of possibly misspecified models in performative prediction, called *plug-in performative optimization*. We show this solution can be far superior to model-agnostic strategies, as long as the misspecification is not too extreme. Our results support the hypothesis that models, even if misspecified, can indeed help with learning in performative settings.

1. Introduction

Predictions have the power to influence the patterns they aim to predict. For example, stock price predictions inform trading decisions and hence prices; traffic predictions influence routing decisions and thus traffic outcomes; recommendations shape users' consumption and thus preferences.

This pervasive phenomenon has been formalized in a framework called *performative prediction* (Perdomo et al., 2020). A central feature that distinguishes the framework from traditional supervised learning is the concept of a *distribution map* $\mathcal{D}(\cdot)$. This object, aimed to capture the feedback from predictions to future observations, is a mapping from predictors f_θ to their induced data distributions $\mathcal{D}(\theta)$. The main goal in performative prediction is thus to deploy a predictor f_θ that will have good performance after deployment, that is, on its induced distribution $\mathcal{D}(\theta)$. Formally, the goal is to choose predictor parameters $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$ so as to minimize the *performative risk*:

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

where $\ell(z;\theta)$ is the loss incurred by predicting on instance z with model θ. Typically, z is a feature–outcome pair (x, y). We refer to

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

as the *performative optimum*.

The main challenge in optimizing the performative risk lies in the fact that the map $\mathcal{D}(\cdot)$ is not known. We only observe samples from $\mathcal{D}(\theta)$ for models θ that have been deployed; we do not observe any feedback for the (typically infinitely many) other models. A key discriminating factor between existing solutions for optimizing under performativity is how they cope with this uncertainty.

One group of methods accounts for the feedback without assuming a problem-specific structure for it. This group includes bandit strategies (Kleinberg et al., 2008; Jagadeesan et al., 2022) and derivative-free optimization (Flaxman et al., 2004; Miller et al., 2021). These methods converge to optima at typically slow—without convexity, even exponentially slow—convergence rates. Moreover, their rates rely on regularity conditions that are out of the learner's control, such as convexity of the performative risk (Miller et al., 2021; Izzo et al., 2021; Dong et al., 2018) or bounded performative effects (Jagadeesan et al., 2022).

A complementary group of methods—an important starting point for this work—takes feedback into account by positing explicit *models* for it. Such models include best-response models for strategic classification (Hardt et al., 2016; Jagadeesan et al., 2021; Levanon & Rosenfeld, 2021; Ghalme et al., 2021), rational-agent models in economics (Spence, 1978; Wooldridge, 2003), and parametric distribution shifts (Izzo et al., 2021; Miller et al., 2021; Izzo et al., 2022),

[1] Department of Statistics, University of California, Berkeley, USA [2] Stanford Data Science and Department of Statistics, Stanford University, USA. Correspondence to: Licong Lin <licong-lin@berkeley.edu>, Tijana Zrnic <tijana.zrnic@stanford.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

among others. To argue that methods building on models find optimal solutions, existing analyses assume that the model is *well-specified*. However, models of social behavior are widely acknowledged to be simplistic representations of real-world dynamics.

Yet, despite the unavoidable misspecification of models, they are ubiquitous in practice. Though their simplicity leads to misspecification, it also allows for efficient, interpretable, and practical solutions. Motivated by this observation, in this work we ask: *can models for performative feedback be useful, even if misspecified?*

1.1. Our Contribution

We initiate a study of the benefits of modeling feedback in performative prediction. We show that models—even if misspecified—can indeed help with learning under performativity.

We begin by defining a general protocol for performative optimization with feedback models, which we call *plug-in performative optimization*. The protocol consists of three steps. First, the learner deploys models $\theta_i \sim \tilde{\mathcal{D}}$ and collects data $z_i \sim \mathcal{D}(\theta_i)$, $i \in [n]$. Here, $\tilde{\mathcal{D}}$ is an exploration distribution of the learner's choosing (for example, it can be uniform on Θ when Θ is bounded). The second step is to use the observations $\{(\theta_i, z_i)\}_{i=1}^n$ to fit an estimate of the distribution map. The map is chosen from a parametric family of possible maps $\boldsymbol{\mathcal{D}}_{\mathcal{B}} = \{\mathcal{D}_\beta\}_{\beta \in \mathcal{B}}$, obtained through modeling. The estimation of the map thus reduces to computing an estimate $\hat{\beta}$. For example, in strategic classification, β could be a parameter quantifying the strategic agents' tradeoff between utility and cost. Finally, the third step is to compute the *plug-in performative optimum*:

$$\hat{\theta}_{\mathrm{PO}} = \underset{\theta\in\Theta}{\arg\min}\, \mathrm{PR}^{\hat{\beta}}(\theta) = \underset{\theta\in\Theta}{\arg\min}\, \mathbb{E}_{z\sim\mathcal{D}_{\hat{\beta}}(\theta)}[\ell(z;\theta)].$$

We prove a general excess-risk bound on $\mathrm{PR}(\hat{\theta}_{\mathrm{PO}}) - \mathrm{PR}(\theta_{\mathrm{PO}})$, showing that the error decomposes into two terms. The first is a *misspecification error term*, $\texttt{MisspecErr}$, which captures the gap between the true performative risk and the plug-in performative risk $\mathrm{PR}^{\hat{\beta}}(\theta)$ in the large-sample regime. This term is irreducible and does not vanish as the sample size n grows. The second is a *statistical error term* that captures the imperfection in fitting $\hat{\beta}$ due to finite samples. For a broad class of problems, our main result can be summarized as follows.

Theorem 1.1 (Informal). *The excess risk of the plug-in performative optimum is bounded by:*

$$\mathrm{PR}(\hat{\theta}_{\mathrm{PO}}) - \mathrm{PR}(\theta_{\mathrm{PO}}) \le c\cdot \texttt{MisspecErr} + \tilde{O}\left(\frac{1}{\sqrt{n}}\right),$$

for some universal constant $c > 0$.

Therefore, although the misspecification error is irreducible, the statistical error vanishes at a *fast rate*. In contrast, model-agnostic strategies such as bandit algorithms (Kleinberg et al., 2008; Jagadeesan et al., 2022) do not suffer from misspecification but have an exceedingly slow, often exponentially slow, statistical rate. For example, the bandit algorithm of (Jagadeesan et al., 2022) has an excess risk of $\tilde{O}(n^{-\frac{1}{d_\theta+1}})$. This is why feedback models are useful—for a finite n, their excess risk can be far smaller than the risk of a model-agnostic strategy due to the rapidly vanishing statistical rate. The statistical rate is fast because it only depends on the parametric estimation rate of $\hat{\beta}$; it does not depend on the complexity of PR.

One important case of performative prediction is *strategic classification*. We apply our general theory to common best-response models in strategic classification. We also conduct numerical evaluations that confirm our theoretical findings. Overall our results support the use of models in optimization under performative feedback.

1.2. Related Work

Performative prediction. We build on the growing body of work studying performative prediction (Perdomo et al., 2020). Existing work studies different variants of retraining (Perdomo et al., 2020; Mendler-Dünner et al., 2020; Drusvyatskiy & Xiao, 2022), which converge to so-called performatively stable solutions, as well as methods for finding performative optima (Miller et al., 2021; Izzo et al., 2021; Jagadeesan et al., 2022). The methods in the latter category are largely model-agnostic and as such converge at slow rates. Exceptions include the study of parametric distribution shifts (Izzo et al., 2021; 2022) and location families (Miller et al., 2021; Jagadeesan et al., 2022), but those analyses crucially rely on the model being well-specified. We are mainly interested in settings where $\mathcal{D}(\theta)$ is a general black-box. Other work in performative prediction includes the study of time-varying shifts (Brown et al., 2022; Izzo et al., 2022; Li & Wai, 2022; Ray et al., 2022), multi-agent settings (Dean et al., 2022; Qiang et al., 2022; Narang et al., 2022; Piliouras & Yu, 2023), causality and robustness (Maheshwari et al., 2022; Mendler-Dünner et al., 2022; Kim & Perdomo, 2023), and it would be valuable to extend our theory on the use of models to those settings.

Strategic classification and economic modeling. Strategic classification (Hardt et al., 2016; Dong et al., 2018; Levanon & Rosenfeld, 2021; Zrnic et al., 2021), as well as other problems studying strategic agent behavior, frequently use models of agent behavior in order to compute Stackelberg equilibria, which are direct analogues of performative optima. However, convergence to Stackelberg equilibria assumes correctness of the models, a challenge we circumvent in this work. We use strategic classification as a primary domain of application of our general theory.

Statistics under model misspecification. Our work is partially inspired by works in statistics studying the benefits and impact of modeling, including under misspecification (White, 1980; 1982; Buja et al., 2019a;b). At a technical level, our results are related to M-estimation (Van der Vaart, 2000; Geer, 2000; Mou et al., 2019) and semiparametric statistics (Tsiatis, 2007; Kennedy, 2022).

Zeroth-order optimization. Plug-in performative optimization serves as an alternative to black-box baselines for zeroth-order optimization, which have previously been studied in performative prediction. These include bandit algorithms (Kleinberg et al., 2008; Jagadeesan et al., 2022) and zeroth-order convex optimization algorithms (Flaxman et al., 2004; Miller et al., 2021). As mentioned, we show that the use of models can give far smaller excess risk, given the fast convergence rates of parametric learning problems.

2. Plug-in Performative Optimization Protocol

We describe the main protocol at the focus of our study and then instantiate it with an example.

We consider the use of parametric models $\boldsymbol{\mathcal{D}}_{\mathcal{B}} := \{\mathcal{D}_\beta\}_{\beta\in\mathcal{B}}$ for modeling the true unknown distribution map $\mathcal{D}$, where $\mathcal{B} \subseteq \mathbb{R}^{d_\beta}$. We denote

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

Since $\boldsymbol{\mathcal{D}}_{\mathcal{B}}$ is a collection of maps, we call it a *distribution atlas*. We emphasize that it need not hold that $\mathcal{D} \in \boldsymbol{\mathcal{D}}_{\mathcal{B}}$; the model could be misspecified.

The protocol for plug-in performative optimization proceeds as follows. First, the learner collects pairs of i.i.d. observations $\{(\theta_i, z_i)\}_{i=1}^n$, where θ_i is deployed according to some exploration distribution $\tilde{\mathcal{D}}$ and $z_i \sim \mathcal{D}(\theta_i)$. The exploration distribution should be "dispersed enough" to enable capturing varied distributions $\mathcal{D}(\theta_i)$ (e.g., uniform, Gaussian with a full-rank covariance, etc). Then, the learner estimates the distribution map by fitting $\hat{\beta}$ based on the collected observations:

$$\hat{\beta} = \widehat{\mathtt{Map}}((\theta_1, z_1), \ldots, (\theta_n, z_n)),$$

where $\widehat{\mathtt{Map}}$ is some model-fitting function. We will consider different criteria for fitting $\hat{\beta}$. We let β^* denote the large-sample limit of $\hat{\beta}$, $\beta^* = \lim_{n\to\infty} \hat{\beta}$. Finally, the learner finds the *plug-in performative optimum*:

$$\hat{\theta}_{\mathrm{PO}} = \operatorname*{arg\,min}_{\theta\in\Theta} \mathrm{PR}^{\hat{\beta}}(\theta) = \operatorname*{arg\,min}_{\theta\in\Theta} \mathbb{E}_{z\sim\mathcal{D}_{\hat{\beta}}(\theta)}[\ell(z;\theta)].$$

We summarize the protocol in Algorithm 1.

Algorithm 1 Plug-in performative optimization

```
Require: distribution atlas $\boldsymbol{\mathcal{D}}_{\mathcal{B}}$, exploration strategy $\tilde{\mathcal{D}}$,
    loss $\ell(z;\theta)$, map-fitting algorithm $\widehat{\mathtt{Map}}$.
1: Deploy $\theta_i \sim \tilde{\mathcal{D}}$, observe $z_i \sim \mathcal{D}(\theta_i)$, $i \in [n]$.
2: Fit distribution map:
    $\hat{\beta} = \widehat{\mathtt{Map}}((\theta_1, z_1), \ldots, (\theta_n, z_n))$, where $\hat{\beta} \in \mathcal{B}$.
3: Compute plug-in performative optimum:
    $\hat{\theta}_{\mathrm{PO}} = \arg\min_{\theta\in\Theta} \mathbb{E}_{z\sim\mathcal{D}_{\hat{\beta}}(\theta)}[\ell(z;\theta)]$.
```

Notice that, since $\mathcal{D}_{\hat{\beta}}$ is known to the learner, we may solve for $\hat{\theta}_{\mathrm{PO}}$ explicitly in Step 3 of the protocol, without collecting any additional real data. In particular, solving for $\hat{\theta}_{\mathrm{PO}}$ incurs only computational complexity—*not* statistical complexity. A detailed discussion on how to execute Step 3 empirically can be found in Appendix C.

A canonical choice of $\widehat{\mathtt{Map}}$ that we will focus on is *empirical risk minimization*:

$$\hat{\beta} = \operatorname*{arg\,min}_{\beta\in\mathcal{B}} \frac{1}{n}\sum_{i=1}^{n} r(\theta_i, z_i; \beta),$$

where r is a loss function. Throughout we will use $\tilde{\theta}$ and $\tilde{z}$ to denote draws $\tilde{\theta} \sim \tilde{\mathcal{D}}, \tilde{z} \sim \mathcal{D}(\tilde{\theta})$. Then, $\beta^* = \arg\min_{\beta\in\mathcal{B}} \mathbb{E}[r(\tilde{\theta}, \tilde{z}; \beta)]$. For example, one can choose $r(\theta, z; \beta) = -\log p_\beta(z;\theta)$ to be the log-likelihood, where $p_\beta(\cdot;\theta)$ is the density under $\mathcal{D}_\beta(\theta)$, in which case $\hat{\beta}$ is the maximum-likelihood estimator. Under this choice,

$$\beta^* = \operatorname*{arg\,max}_{\beta\in\mathcal{B}} \mathbb{E}[\log p_\beta(\tilde{z};\tilde{\theta})] = \operatorname*{arg\,min}_{\beta\in\mathcal{B}} \mathrm{KL}(\bar{\mathcal{D}}, \bar{\mathcal{D}}_\beta).$$

Here, $\bar{\mathcal{D}}$ is the distribution of $\tilde{z}$, that is, the distribution map $\mathcal{D}(\theta)$ averaged over $\theta \sim \tilde{\mathcal{D}}$. We similarly define $\bar{\mathcal{D}}_\beta$. Therefore, β^* is the KL projection of the true data-generating process onto the considered distribution atlas.

Example 1 (Biased coin flip)**.** *To build intuition for the introduced concepts, we consider an illustrative example. Suppose we want to predict the outcome of a biased coin flip, where the bias arises due to performative effects. The outcome is generated as* $z \sim \mathcal{D}(\theta) = \mathrm{Bern}(0.5 + \mu\theta + \eta\theta^2)$*, where* $\mu \in (0, 0.5), \eta \in (0, 0.5 - \mu)$*. The parameter* $\theta \in [0, 1]$ *aims to predict the outcome while minimizing the squared loss,* $\ell(z;\theta) = (z - \theta)^2$*. Suppose that we know that* θ *introduces a bias to the coin flip, but we do not know how strongly or in what way. We thus choose a simple model for the bias,* $\mathcal{D}_\beta(\theta) = \mathrm{Bern}(0.5 + \beta\theta)$*, and fit* β *in a data-driven way. To do so, we deploy* $\theta_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}[0, 1]$ *and observe* $z_i \sim \mathcal{D}(\theta_i)$*, for* $i \in [n]$*. One natural way to fit the distribution map is to solve*

$$\hat{\beta} = \operatorname*{arg\,min}_{\beta} \frac{1}{n}\sum_{i=1}^{n}(z_i - 0.5 - \beta\theta_i)^2.$$

Finally, we compute the plug-in performative optimum as

$$\hat{\theta}_{\mathrm{PO}} = \operatorname*{arg\,min}_{\theta} \mathbb{E}_{z\sim\mathcal{D}_{\hat{\beta}}(\theta)}[(z - \theta)^2] = \frac{1 - \hat{\beta}}{2 - 4\hat{\beta}}.$$

It is not hard to show that the population limit of $\hat{\beta}$ is equal to $\beta^* = \mu + 0.75\eta$. Therefore, if the feedback model is well-specified, meaning $\eta = 0$, then β^* indeed recovers the true distribution map, and $\hat{\theta}_{\mathrm{PO}}$ converges to the true performative optimum.

3. Excess Risk

We study the excess risk of plug-in performative optimization. The key takeaway of this section is that the excess risk depends on two sources of error: one is the *misspecification error* due to the fact that, often, $\mathcal{D} \notin \boldsymbol{\mathcal{D}}_{\mathcal{B}}$; the other is the *statistical error* due to the gap between $\hat{\beta}$ and β^*.

Formally, define

$$\begin{aligned}\texttt{MisspecErr} &= \sup_{\theta\in\Theta} |\mathrm{PR}^{\beta^*}(\theta) - \mathrm{PR}(\theta)|, \\ \texttt{StatErr}_n &= \sup_{\theta\in\Theta} |\mathrm{PR}^{\beta^*}(\theta) - \mathrm{PR}^{\hat{\beta}}(\theta)|.\end{aligned}$$

We note that the statistical error depends on the sample size n, while the misspecification error is irreducible even in the large-sample limit. In later sections we will show that the statistical error vanishes at a *fast rate*, namely $\tilde{O}\left(\frac{1}{\sqrt{n}}\right)$, for a broad class of problems. In Theorem 3.1 we state a general bound on the excess risk in terms of these two sources of error.

Theorem 3.1. *The excess risk of the plug-in performative optimum is bounded by:*

$$\mathrm{PR}(\hat{\theta}_{\mathrm{PO}}) - \mathrm{PR}(\theta_{\mathrm{PO}}) \leq 2\left(\texttt{MisspecErr} + \texttt{StatErr}_n\right).$$

Theorem 3.1 illuminates the benefits of feedback models: if the model is a reasonable approximation, the misspecification error is not too large; at the same time, due to the parametric specification of the distribution atlas, the statistical error vanishes quickly. Therefore, we conclude that *even misspecified* models can lead to lower excess risk than entirely model-agnostic strategies such as bandit algorithms.

Remark 3.2. It should be noted that there may be numerical inaccuracy in solving for $\hat{\theta}_{\mathrm{PO}}$ in Step 3 of Algorithm 1. However, the bound of Theorem 3.1 degrades smoothly: if a δ-suboptimal solution is attained, then the excess risk increases by at most δ. As mentioned before, δ is not dependent on n; it only depends on the amount of computation.

In the rest of this section we give fine-grained bounds on the misspecification error and the statistical error under appropriate regularity assumptions, providing intuition via examples along the way. The most natural way to bound the misspecification error is in terms of a distributional distance between the true distribution map $\mathcal{D}(\theta)$ and the modeled distribution map $\mathcal{D}_{\beta^*}(\theta)$. We define the misspecification of a distribution atlas.

Definition 3.3 (Misspecification)**.** We say that a distribution atlas is η-misspecified in distance $\texttt{dist}$ if, for all $\theta \in \Theta$, it holds that $\texttt{dist}(\mathcal{D}_{\beta^*}(\theta), \mathcal{D}(\theta)) \leq \eta$.

We will measure misspecification in either total-variation distance or Wasserstein (i.e. earth mover's) distance. Depending on the problem setting, one of the two distances will yield a smaller misspecification parameter and thus a tighter rate according to Theorem 3.1.

We will also require that the atlas is "smooth" in the analyzed distance.

Definition 3.4 (Smoothness)**.** We say that a distribution atlas is ϵ-smooth in distance $\texttt{dist}$ if, for all $\beta, \beta' \in \mathcal{B}$ and $\theta \in \Theta$, it holds that $\texttt{dist}(\mathcal{D}_{\beta}(\theta), \mathcal{D}_{\beta'}(\theta)) \leq \epsilon\|\beta - \beta'\|$.

Unless stated otherwise, $\|\cdot\|$ denotes the ℓ_2-norm. In some examples the parameter of the atlas will be a matrix, in which case the norm will be the operator norm $\|\cdot\|_{\mathrm{op}}$. It is important to note that, while the misspecification parameter is a property of the true distribution map, the smoothness parameter is entirely in the learner's control, as it is solely a property of the chosen distribution atlas.

In what follows, Section 3.1 and Section 3.2 focus on bounding the misspecification error. Section 3.3 focuses on bounding the statistical error with an explicit rate.

3.1. Total-Variation Misspecification

First we consider misspecification in total-variation (TV) distance. Building on Theorem 3.1, we obtain the following excess-risk bound as a function of TV misspecification.

Corollary 3.5. *Suppose the distribution atlas is η_{TV}-misspecified and ϵ_{TV}-smooth in TV distance. Moreover, suppose that $|\ell(z;\theta)| \leq B_\ell$ and $\|\hat{\beta} - \beta^*\| \leq C_n$. Then, the excess risk of the plug-in performative optimum is bounded by:*

$$\mathrm{PR}(\hat{\theta}_{\mathrm{PO}}) - \mathrm{PR}(\theta_{\mathrm{PO}}) \leq 4B_\ell \cdot \eta_{\mathrm{TV}} + 4B_\ell \cdot \epsilon_{\mathrm{TV}} \cdot C_n.$$

Corollary 3.5 shows that plug-in performative optimization is efficient as long as the distribution atlas is smooth, not too misspecified, and the rate of estimation C_n is fast. In Section 3.3 we will prove convergence rates C_n when $\hat{\beta}$ is a sufficiently regular empirical risk minimizer.

We now build intuition for the relevant terms in Corollary 3.5 through examples. First we give a couple of examples of distribution atlases and bound their smoothness parameter ϵ_{TV}.

Example 2 (Mixture model)**.** *Suppose that we have k candidate distribution maps $\{\mathcal{D}^{(i)}(\theta)\}_{i=1}^k$. We would like to find a combination of these maps that approximates the true map $\mathcal{D}(\theta)$ as closely as possible. To do so, we can define $\mathcal{D}_\beta(\theta) = \sum_{i=1}^k \beta_i \mathcal{D}^{(i)}(\theta)$, where $\beta \in [0,1]^k$ defines the mixture weights. This model is smooth in TV distance: $\mathrm{TV}(\mathcal{D}_\beta(\theta), \mathcal{D}_{\beta'}(\theta)) \leq \frac{1}{2}\|\beta - \beta'\|_1 \leq \frac{\sqrt{k}}{2}\|\beta - \beta'\|_2$.*

Example 3 (Self-fulfilling prophecy)**.** *Suppose that we want to model outcomes that follow a "self-fulfilling prophecy," meaning that predicting a certain outcome makes it more likely for the outcome to occur. Assume we have historical data of feature–label pairs before the model was deployed. Denote the resulting empirical distribution by* $\mathcal{D}_0 = \mathcal{D}_0^X \times \mathcal{D}_0^{Y|X}$*. We assume the features are non-performative; only the labels exhibit performative effects. Then, we can model the label distribution as* $\mathcal{D}_\beta^{Y|X}(\theta) = (1-\beta)\mathcal{D}_0^{Y|X} + \beta\delta_{f_\theta(X)}$*, where* $\delta_{f_\theta(X)}$ *denotes a point mass at the predicted label. Here,* $\beta \in [0,1]$ *tunes the strength of performativity:* $\beta = 0$ *implies no performativity, while* $\beta = 1$ *a perfect self-fulfilling prophecy. This atlas has TV-smoothness equal to* $\epsilon_{\mathrm{TV}} = 1$*.*

Next, we describe a general type of misspecification that implies a bound on η_{TV}.

Example 4 ("Typically" well-specified model)**.** *Suppose that the data distribution consists of observations about strategic individuals. Suppose that a* $(1-p)$*-fraction of the population is "rational" and acts in a predictable fashion. The remaining* p*-fraction acts arbitrarily. Then, if we model the predictable behavior appropriately, meaning* $\mathcal{D}_{\beta^*}(\theta)$ *follows the distribution produced by the rational agents, the misspecification parameter* η_{TV} *is at most* p*. More generally, if we have* $\mathcal{D}(\theta) = (1-p)\mathcal{D}_{\beta^*}(\theta) + p\tilde{\mathcal{D}}(\theta)$*, where* $\tilde{\mathcal{D}}(\theta)$ *is an arbitrary component, then* $\eta_{\mathrm{TV}} \leq p$*.*

3.2. Wasserstein Misspecification

We next consider misspecification in Wasserstein (i.e. earth mover's) distance. Building on Theorem 3.1, we bound the excess-risk via Wasserstein misspecification.

Corollary 3.6. *Suppose that the distribution atlas is* η_W*-misspecified and* ϵ_W*-smooth in Wasserstein distance. Moreover, suppose that the loss* $\ell(z;\theta)$ *is* L_z*-Lipschitz in* z*, and that* $\|\hat{\beta} - \beta^*\| \leq C_n$*. Then, the excess risk of the plug-in performative optimum is bounded by:*

$$\mathrm{PR}(\hat{\theta}_{\mathrm{PO}}) - \mathrm{PR}(\theta_{\mathrm{PO}}) \leq 2L_z \cdot \eta_W + 2L_z \cdot \epsilon_W \cdot C_n.$$

As in Corollary 3.5, we see that the excess risk of the plug-in performative optimum is small as long as the distribution atlas is smooth, not overly misspecified, and the rate C_n is sufficiently fast.

Below we give an example of a natural distribution atlas and characterize its Wasserstein smoothness.

Example 5 (Performative outcomes)**.** *As in Example 3, suppose that we have data of feature–outcome pairs before any model deployment, and suppose that only the outcomes are performative while the features are static. Let* $\mathcal{D}_0$ *denote the historical data distribution. We assume that a predictor* θ *affects the outcomes only through its predictions* $f_\theta(x)$*. One simple way to model such feedback is via an additive effect on the outcomes. Formally, we define* $(x,y) \sim \mathcal{D}_\beta(\theta) \Leftrightarrow (x_0,y_0) \sim \mathcal{D}_0, x = x_0, y = y_0 + \beta \cdot f_\theta(x)$*. As in Example 3,* $\beta \in \mathbb{R}$ *controls the strength of performativity. This atlas is* ϵ_W*-smooth in Wasserstein distance for* $\epsilon_W = \sup_\theta \mathbb{E}_{x \sim \mathcal{D}_0^X}[|f_\theta(x)|]$*.*

One way that misspecification can arise is due to *omitted-variable bias*. We illustrate this in the following example and explicitly bound the misspecification parameter η_W.

Example 6 (Omitted-variable bias)**.** *Suppose that only a subset of the coordinates* $\mathcal{I} \subseteq [d]$ *of* θ *induce performative effects. This can happen in linear or logistic regression, where the coordinates of* θ *measure feature importance, but only a subset of the features are manipulable. Specifically, assume the data follows a* location family *model:* $z \sim \mathcal{D}(\theta) \Leftrightarrow z = z_0 + \tilde{\mathcal{M}}\theta_{\mathcal{I}}$*, where* $\tilde{\mathcal{M}} \in \mathbb{R}^{d_z \times |\mathcal{I}|}$ *is a true parameter of the shift and* z_0 *is a zero-mean draw from a base distribution* $\mathcal{D}_0$*. Suppose the model omits one performative coordinate by mistake:* $z \sim \mathcal{D}_{\mathcal{M}}(\theta) \Leftrightarrow z = z_0 + \mathcal{M}\theta_{\mathcal{I}'}$*, where* $\mathcal{I}' = \mathcal{I} \setminus \{i_{\mathrm{miss}}\}$ *and* $\mathcal{M} \in \mathbb{R}^{d_z \times |\mathcal{I}'|}$*. If* $\widehat{\mathcal{M}}$ *is fit via least-squares,* $\widehat{\mathcal{M}} = \arg\min_{\mathcal{M}} \frac{1}{n}\sum_{i=1}^n \|z_i - \mathcal{M}\theta_{i,\mathcal{I}'}\|^2$*, and* $\tilde{\mathcal{D}}$ *is a product distribution, then the population-level counterpart of* $\widehat{\mathcal{M}}$ *is equal to* $\mathcal{M}^* = \tilde{\mathcal{M}}_{\mathcal{I}'}$*, which denotes the restriction of* $\tilde{\mathcal{M}}$ *to the columns indexed by* $\mathcal{I}'$*. Putting everything together, we can conclude that the misspecification due to the omitted coordinate is* $\eta_W = \sup_\theta \mathcal{W}(\mathcal{D}(\theta), \mathcal{D}_{\mathcal{M}^*}(\theta)) \leq B\|\tilde{\mathcal{M}}_{i_{\mathrm{miss}}}\|$*, where we assume the* i_{miss}*-coordinate of* $|\theta|$ *is at most* B*.*

3.3. Bounding the Estimation Error

We saw that the statistical error is driven by the estimation rate of β^*. We show that for a broad class of problems the rate is $\tilde{O}(n^{-\frac{1}{2}})$. We focus on map-fitting via *empirical risk minimization* (ERM):

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{B}} \frac{1}{n}\sum_{i=1}^n r(\theta_i, z_i; \beta), \tag{1}$$

where $\mathcal{B} \subset \mathbb{R}^{d_\beta}$ is bounded and convex. We denote $r(\beta) = \mathbb{E}[r(\tilde{\theta}, \tilde{z}; \beta)]$, and note that $\beta^* = \arg\min_{\beta \in \mathcal{B}} r(\beta)$. To establish the error bound, we rely on the following assumptions.

Assumption 3.7. ERM problem (1) is regular if:

(a) $r(\beta)$ is convex and μ-strongly convex at β^*, and the Hessian $\nabla^2 r(\beta)$ is σ_r-Lipschitz;

(b) $\forall \beta \in \mathcal{B}$, the gradient $\nabla r(\tilde{\theta}, \tilde{z}; \beta)$ is a subexponential vector with parameter $B_r > 0$;

(c) $\forall u, u' \in \mathcal{S}^{d_\beta - 1}$, $\sup_{\beta \in \mathcal{B}} u^\top \nabla^2 r(\tilde{\theta}, \tilde{z}; \beta) u'$ is subexponential with parameter $L_r > 0$.

We will see that Assumption 3.7 is satisfied in many important settings. Condition (a) is fairly standard; conditions (b,c) seem less standard because we want to make them realistic for problems with diverging dimension. They immediately hold if $\nabla r(\theta, z; \beta), \nabla^2 r(\theta, z; \beta)$ are bounded, which are standard conditions.

The following lemma is the key tool for obtaining our main excess-risk bounds with a *fast* rate, stated in Theorem 3.9.

Lemma 3.8. *Assume the map-fitting algorithm is regular (Ass. 3.7) and that $\frac{n}{\log n} \geq C(d_\beta + \log(1/\delta))$ for a sufficiently large $C > 0$. Then, for some constant $C' > 0$, with probability $1 - \delta$ it holds that*

$$\|\hat{\beta} - \beta^*\| \leq C' \sqrt{\log n} \sqrt{\frac{d_\beta + \log(1/\delta)}{n}}.$$

Theorem 3.9. *Assume the map-fitting algorithm is regular (Ass. 3.7) and that $\frac{n}{\log n} \geq C(d_\beta + \log(1/\delta))$ for a sufficiently large $C > 0$.*

- *If the distribution atlas is η_{TV}-misspecified and ϵ_{TV}-smooth in total-variation distance, and $|\ell(z;\theta)| \leq B_\ell$, then there exists a $C' > 0$ such that, with probability $1 - \delta$:*

$$\mathrm{PR}(\hat{\theta}_{\mathrm{PO}}) - \mathrm{PR}(\theta_{\mathrm{PO}}) \leq 4B_\ell \eta_{\mathrm{TV}} + C' B_\ell \epsilon_{\mathrm{TV}} \sqrt{\log n} \sqrt{\frac{d_\beta + \log(1/\delta)}{n}}.$$

- *If the distribution atlas is η_W-misspecified and ϵ_W-smooth in Wasserstein distance, and $\ell(z;\theta)$ is L_z-Lipschitz in z, then there exists a $C' > 0$ such that, with probability $1 - \delta$:*

$$\mathrm{PR}(\hat{\theta}_{\mathrm{PO}}) - \mathrm{PR}(\theta_{\mathrm{PO}}) \leq 2L_z \eta_W + C' L_z \epsilon_W \sqrt{\log n} \sqrt{\frac{d_\beta + \log(1/\delta)}{n}}.$$

The excess risk is thus bounded by the sum of a term due to misspecification and a *fast* statistical rate. To contrast this with a model-agnostic rate, the bandit algorithm of (Jagadeesan et al., 2022) does not suffer from misspecification but has an *exponentially* slow excess risk, $\tilde{O}(n^{-1/(d_\theta+1)})$.

The analysis underlying Theorem 3.9 is based on standard ERM analyses under model misspecification, though there are differences. The main one is that our setting requires that we analyze the error in terms of the difference between parameters $\hat{\beta} - \beta^*$, as opposed to the excess risk $r(\hat{\beta}) - r(\beta^*)$ in the standard ERM analysis. This difference requires new tools and assumptions akin to those in Assumption 3.7.

4. Applications

We apply our theory and plug-in performative optimization to several problems with performative feedback, building on prevalent models for those problems. In each problem, we prove the model's smoothness and fast estimation of β^*.

4.1. Strategic Regression

We begin by considering strategic regression. Here, a population of individuals described by (x, y) strategically responds to a deployed predictor f_θ. For example, the predictor could be $f_\theta(x) = \theta^\top x$.

Distribution atlas. The strategic responses consist of manipulating features in order to maximize a utility function, which is often equal to the prediction itself. Formally, given an individual with features x_0, a commonly studied response model is $g_\beta(x_0, \theta) = \arg\max_x (u_\theta(x) - \frac{1}{2\beta}\|x - x_0\|^2)$, where $u_\theta(x)$ is a concave utility function and the second term captures the cost of feature manipulations. Here, $\beta \in [\beta_{\min}, \beta_{\max}] \subseteq \mathbb{R}_+$ trades off utility and cost. The natural distribution atlas capturing the above response model is obtained as follows. Suppose that we have a historical distribution of feature–label pairs $\mathcal{D}_0$. Then, let:

$$(x, y) \sim \mathcal{D}_\beta(\theta) \Leftrightarrow (x_0, y) \sim \mathcal{D}_0, x = g_\beta(x_0, \theta). \quad (2)$$

Claim 4.1. *If $\|\nabla u_\theta(x)\| \leq B_u$ and $\nabla u_\theta(x)$ is L_u-Lipschitz, then $\{\mathcal{D}_\beta\}_\beta$ has*

$$\epsilon_W \leq \frac{B_u}{1 - \beta_{\max} L_u}.$$

This bound is attained for $u_\theta(x) = x^\top \theta$. There, $L_u = 0, B_u = \sup_{\theta\in\Theta} \|\theta\|$, so the bound equals $\epsilon_W \leq \max_{\theta\in\Theta} \|\theta\|$. This is tight because $\sup_{x,\theta\in\Theta} \|g_\beta(x,\theta) - g_{\beta'}(x,\theta)\| = |\beta - \beta'| \sup_{\theta\in\Theta} \|\theta\|$.

Map fitting. We can fit $\hat{\beta}$ via maximum-likelihood estimation (MLE). Suppose that $\mathcal{D}_0^X$ has a density and denote it by p_0^X. Then, we can let

$$\hat{\beta} = \arg\min_{\beta\in\mathcal{B}} -\frac{1}{n} \sum_{i=1}^n \log\left(p_0^X(x_i - \beta \nabla u_{\theta_i}(x_i))\right).$$

Given the first-order optimality condition for $g_\beta(x_0, \theta)$, under mild regularity and correct specification of the model, meaning $\mathcal{D}(\theta) = \mathcal{D}_\beta(\theta)$ for some β, this map-fitting strategy ensures $\eta_W = 0$, as expected. For example, if $\mathcal{D}_0^X = \mathcal{N}(0, \sigma^2 I)$, MLE reduces to

$$\hat{\beta} = \arg\min_{\beta\in\mathcal{B}} \frac{1}{n} \sum_{i=1}^n \|x_i - \beta \nabla u_{\theta_i}(x_i)\|^2.$$

Least-squares makes sense even if the features are not Gaussian; it just coincides with MLE for Gaussians. We show a fast estimation rate of least-squares under mild conditions via Lemma 3.8.

Claim 4.2. *If $\mathbb{E}[\|\nabla u_{\tilde{\theta}}(\tilde{x})\|^2] > 0$, $\tilde{x}$ and $\nabla u_{\tilde{\theta}}(\tilde{x})$ are subgaussian, and $\frac{n}{\log n} \geq C(1 + \log(1/\delta))$ for a sufficiently large $C > 0$, then*

$$|\hat{\beta} - \beta^*| \leq C' \sqrt{\log n} \sqrt{\frac{1 + \log(1/\delta)}{n}}$$

with probability $1 - \delta$.

If, in addition, the loss function $\ell(z; \theta)$ is L_z-Lipschitz in z, combining Claim 4.1, Claim 4.2, and Corollary 3.6 gives an upper bound on the excess risk $\mathrm{PR}(\hat{\theta}_{\mathrm{PO}}) - \mathrm{PR}(\theta_{\mathrm{PO}})$.

4.2. Binary Strategic Classification

Next, we consider binary strategic classification, in which a population of strategic individuals described by (x, y) takes strategic actions in order to reach a decision boundary. We assume the learner's decision rule is obtained by thresholding a linear model, $f_\theta(x) = \mathbf{1}\{\theta^\top x \geq T\}$, for some T. Without loss of generality we assume $\|\theta\| = 1$, since the rule is invariant to rescaling θ and T.

Distribution atlas. A common model assumes that the individuals have a budget $\beta > 0$ on how much they can change their features (Kleinberg & Raghavan, 2020; Chen et al., 2020; Zrnic et al., 2021). The individuals move to the decision boundary if it is within ℓ_2 distance β. Formally, an individual with features x_0 responds with $g_\beta(x_0, \theta) = x_0 + \theta(T - x_0^\top \theta)$ if $x_0^\top \theta \in [T - \beta, T)$, and does not move otherwise. The natural distribution atlas corresponding to the above model is defined as in Eq. (2), for a given base distribution $\mathcal{D}_0$. We show that this atlas is smooth in total-variation distance.

Claim 4.3. *If, for all θ, $x_0^\top \theta$ has a density upper bounded by ϕ_u, then $\{\mathcal{D}_\beta\}_\beta$ has $\epsilon_{\mathrm{TV}} \leq \phi_u$.*

Map fitting. According to the atlas, all individuals with $x_0^\top \theta \in [T - \beta, T)$ move to the decision boundary, defined by $x^\top \theta = T$. Therefore, one can estimate the individuals' budget by finding $\hat{\beta}$ such that $\mathbb{P}\{x_0^\top \tilde{\theta} \in [T - \hat{\beta}, T]\} = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{x_i^\top \theta_i \in (T \pm \epsilon)\}$, for a small $\epsilon > 0$. The latter term estimates the mass in a small neighborhood of the boundary. Therefore, $\mathbb{P}\{x_0^\top \tilde{\theta} \in [T - \beta^*, T]\} = \mathbb{P}\{\tilde{x}^\top \tilde{\theta} \in (T \pm \epsilon)\}$. For simplicity we assume $x_0^\top \tilde{\theta}$ has a density on $\mathbb{R}$ so that $\hat{\beta}$ exists. Note that $\mathbb{P}\{x_0^\top \tilde{\theta} \in [T - \beta, T]\}$ is known for all β because it is a property of the base distribution, so finding $\hat{\beta}$ reduces to estimating $\mathbb{P}\{\tilde{x}^\top \tilde{\theta} \in (T \pm \epsilon)\}$.

Claim 4.4. *If $x_0^\top \tilde{\theta}$ has a density lower bounded by ϕ_l, then*

$$|\hat{\beta} - \beta^*| \leq \frac{1}{\phi_l} \sqrt{\frac{\log(2/\delta)}{2n}}$$

with probability $1 - \delta$.

If the learner's loss is bounded, putting together Claim 4.3, Claim 4.4, and Corollary 3.5 gives an upper bound on the excess risk $\mathrm{PR}(\hat{\theta}_{\mathrm{PO}}) - \mathrm{PR}(\theta_{\mathrm{PO}})$.

4.3. Location Families

Lastly, we consider general location families (Miller et al., 2021; Jagadeesan et al., 2022; Ray et al., 2022), in which the deployment of θ leads to performativity via a linear shift. This model often appears in strategic classification with linear or logistic regression, and can capture performativity only in certain features (Miller et al., 2021).

Distribution atlas. The location-family model is defined by $z \sim \mathcal{D}_{\mathcal{M}}(\theta) \Leftrightarrow z = \mathcal{M}\theta + z_0$, where $\mathcal{M} \in \mathbb{R}^{d_z \times d_\theta}$ is a matrix that parameterizes the shift, and z_0 is a sample from a zero-mean base distribution $\mathcal{D}_0$. We assume $\sup_{\theta \in \Theta} \|\theta\| \leq B_\theta$. It is not hard to see that the atlas is smooth in $\|\cdot\|_{\mathrm{op}}$.

Claim 4.5. *The atlas $\{\mathcal{D}_{\mathcal{M}}\}_{\mathcal{M}}$ has $\epsilon_W \leq B_\theta$.*

Map fitting. We fit the distribution map via least-squares:

$$\widehat{\mathcal{M}} = \arg\min_{\mathcal{M}} \frac{1}{n} \sum_{i=1}^n \|z_i - \mathcal{M}\theta_i\|^2.$$

Thus, $\mathcal{M}^* = \arg\min_{\mathcal{M}} \mathbb{E}[\|\tilde{z} - \mathcal{M}\tilde{\theta}\|^2]$. We provide control on the estimation error below.

Claim 4.6. *Assume $\tilde{\mathcal{D}}$ is zero-mean and subgaussian with $\kappa_{\min} I \preceq \mathbb{E}[\tilde{\theta}\tilde{\theta}^\top] \preceq \kappa_{\max} I$. Further, for all $u \sim \mathcal{S}^{d_\theta - 1}, v \sim \mathcal{S}^{d_z - 1}$, assume $u^\top \tilde{\theta} \cdot v^\top \tilde{z}$ is subexponential with parameter $L_{\theta z}$ and $\|\mathbb{E}[\tilde{\theta}\tilde{z}^T]\|_{\mathrm{op}} \leq B$. Then, if $n \geq C(d_\theta + d_z + \log(1/\delta))$ for some sufficiently large $C > 0$, there exists $C' > 0$ such that with probability $1 - \delta$ we have*

$$\|\widehat{\mathcal{M}} - \mathcal{M}^*\|_{\mathrm{op}} \leq C' \sqrt{\frac{d_\theta + d_z + \log(1/\delta)}{n}}.$$

The above assumptions hold under mild regularity conditions when the model is nearly well-specified. If the loss is L_z-Lipschitz in z, then using Claim 4.5, Claim 4.6, and Corollary 3.6, we can bound the excess performative risk $\mathrm{PR}(\hat{\theta}_{\mathrm{PO}}) - \mathrm{PR}(\theta_{\mathrm{PO}})$.

5. Experiments

We confirm the qualitative takeaways of our theory empirically. We compare our approach with two model-agnostic algorithms: derivative-free optimization (DFO) (Flaxman et al., 2004) and greedy SGD (Mendler-Dünner et al., 2020). The latter naively retrains while ignoring the feedback; it is a practical heuristic but only converges to stable points. For the location-family experiment, we also compare our approach with PerfGD (Izzo et al., 2021), which approximates the performative gradient via numerical methods. We refer the reader to Appendix B for further details.

5.1. Location Family

We start with the location-family setting. We assume the true map follows a linear model with a quadratic term, $z_i = b +$

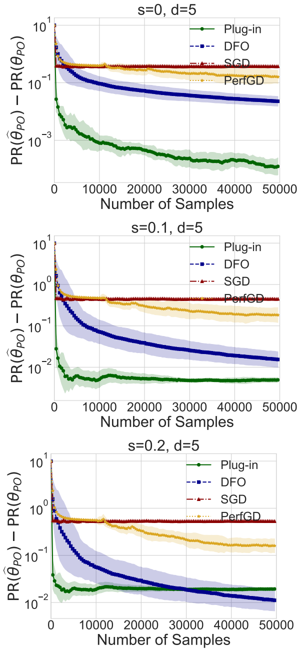

Figure 1. (Location family) Excess risk of plug-in performative optimization, DFO, greedy SGD, and PerfGD with ±1 standard deviation on a logarithmic scale.

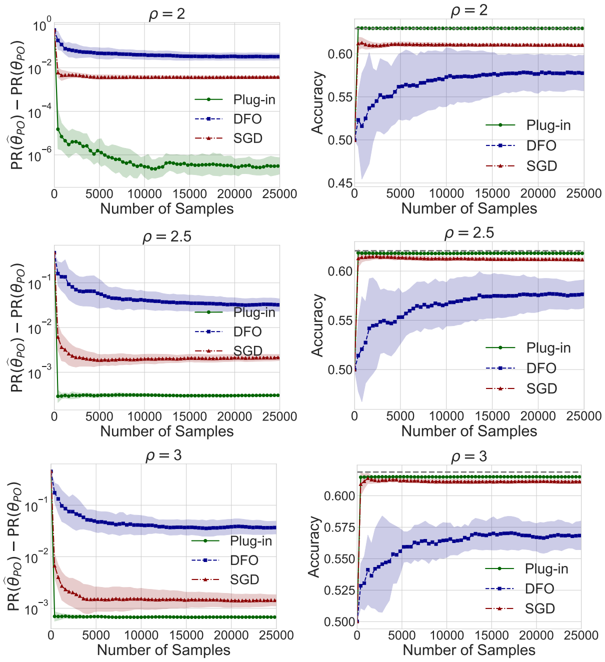

Figure 2. (Strategic regression) Excess risk and accuracy of plug-in performative optimization, DFO, and greedy SGD, with ±1 standard deviation on a logarithmic scale.

$\mathcal{M}_1\theta_i + s\mathcal{M}_2(\theta_i \circ \theta_i) + z_{0,i}$, where $z_i, \theta_i, b \in \mathbb{R}^d$, $\theta_i \circ \theta_i := (\theta_{i,1}^2, \dots, \theta_{i,d}^2)$ represents the quadratic effect, and $z_{0,i} \sim \mathcal{N}(0, \sigma^2 I_d)$. The parameter $s \geq 0$ varies the magnitude of misspecification. We want to minimize the loss $\ell(z;\theta) = \|z - \theta\|^2$, and we use a simple linear model to approximate $\mathcal{D}(\theta)$, i.e., $z \sim \mathcal{D}_{\mathcal{M}}(\theta) \Leftrightarrow z \overset{d}{=} b + \mathcal{M}\theta + \ z_0$. To fit $\widehat{\mathcal{M}}$, we use the loss $r(\theta, z; \mathcal{M}) = \|z - \mathcal{M}\theta\|^2$. We vary $d \in \{5, 10\}$ and let $\mathcal{M}_i = \tilde{\mathcal{M}}_i / \|\tilde{\mathcal{M}}_i\|_{\text{op}}$, $i \in \{1, 2\}$ where $\tilde{\mathcal{M}}_i \in \mathbb{R}^{d \times d}$ have entries generated i.i.d. from $\mathcal{N}(0, 1)$. We let $b \sim \mathcal{N}(0, I_d)$, $\sigma = 0.5$, and $\Theta = \{\theta : \|\theta\| \leq 1\}$.

In Figure 1 we see that the excess risk of our algorithm converges rapidly to a value that reflects the degree of misspecification. It approaches zero for $s = 0$ (top panel) due to no misspecification and stabilizes at a nonzero value for $s > 0$ (middle and bottom panels), consistent with our theory. In contrast, the risk of both PerfGD and DFO reduce slowly, while SGD quickly reaches a suboptimal value.

5.2. Strategic Regression

We next consider strategic regression. We first generate 5000 i.i.d. base samples (x_i, y_i) where $x_i \sim \mathcal{N}(0, I_{d_x})$, and $y_i \in \{0, 1\}$ follows from a logistic model with a fixed parameter vector $\eta \sim \mathrm{Unif}(\mathcal{S}^{d_x - 1})$. Denote the joint empirical distribution of (x_i, y_i) by $\mathcal{D}_0$. The true distribution map, $(x, y) \sim \mathcal{D}(\theta)$, is defined via

$$(x_0, y_0) \sim \mathcal{D}_0, y = y_0, x = \arg\max_{x' \in \mathbb{R}^{d_x}} (\theta^\top x' - \frac{1}{2\tilde{\beta}} \|x' - x_0\|_\rho^\rho),$$

for some $\rho > 1$. We set $d_x = 5$ and $\tilde{\beta} = 2$. To construct a model for $\mathcal{D}(\cdot)$, we follow the procedure from Section 4.1, using the linear utility and quadratic cost. Thus, $\rho = 2$ results in correct specification. We choose $\ell(z;\theta)$ to be the logistic loss with a ridge penalty and $\Theta = \{\theta : \|\theta\| \leq 1\}$.

We examine the well-specified scenario $\rho = 2$ and two misspecified cases $\rho \in \{2.5, 3\}$, comparing our method with DFO and greedy SGD in Figure 2. We see our algorithm quickly tends to zero excess risk when $\rho = 2$ (top left). When $\rho \neq 2$ (middle and bottom left), the algorithm still converges, albeit to a suboptimal point. SGD converges to a suboptimal point with nonzero excess risk, while the excess risk of the DFO method decreases at a slow rate. We remark that PerfGD is not applicable in our strategic regression setting as $\mathcal{D}(\theta)$ does not admit a probability density function. Similar trends appear for test accuracy.

6. Discussion

We discuss several limitations and comment on a few technical aspects of our work. Along the way, we discuss valuable future directions in the context of modeling the distribution map in performative prediction.

First, our approach may be of limited use when the learner has very little prior knowledge of the true data-generating process, as this prevents them from choosing an atlas in an informed way. This is a hard scenario, and it is possible that no existing method would lead to a satisfactory performance.

For example, while the model-free approach of Jagadeesan et al. (2022) is asymptotically more robust, for a reasonable number of samples it is not clear that it would outperform our approach with a highly misspecified model. One valuable future direction is to choose a family of distribution maps $\boldsymbol{\mathcal{D}}_{\mathcal{B}}$ that is sufficiently complex and has strong expressive power, e.g., neural networks, in cases where the learner is very uncertain about the true map. When the number of samples is suitably large, this may give a small excess risk as the misspecification error is smaller with a more expressive model. While in the current paper we mainly focus on simple models with closed-form expressions, we are hopeful that our approach could be valuable in such settings with black-box modeling as well.

Modeling could be particularly useful if one considers the fact that in reality data distributions take time to shift after model deployment and do not generate i.i.d. observations. Such time-varying shifts were, in fact, modeled in prior work (Brown et al., 2022; Izzo et al., 2022; Li & Wai, 2022; Ray et al., 2022). It would be interesting to study the impacts of modeling the time-varying aspect of distribution maps.

Many of our examples assumed access to historical data. However, this assumption is not fundamental to our framework. Often this just means that there was another historical model in place under which the data was collected. We can model this as "step 0" in our framework, where we first deploy a "default" model θ_0 (e.g. $\theta_0 = 0$) and collect data points drawn from $\mathcal{D}(\theta_0) \equiv \mathcal{D}_0$ (for instance, in Example 5 this would correspond to deploying a model with $f_\theta(x) = 0$ for all x; there exist analogues for other problems as well). This alternative view would simply incur an additional statistical error term due to having finite-sample access to $\mathcal{D}(\theta_0)$.

In this work, we used standard ERM to estimate the distribution map. However, the learner may want to incorporate criteria other than model fit (which is the focus on ERM) in this process. For example, the learner may want to use a classical model selection criterion such as AIC or BIC (see, e.g., Hastie et al. (2009)) to regularize towards "simpler" models and run ERM afterwards. Investigating model selection criteria beyond ERM for distribution map estimation is another valuable future direction.

Impact Statement

We have analyzed performative prediction with misspecified models. Our results highlight the statistical gains of using models, however modeling has consequences far beyond statistical efficiency. On the positive side, modeling can help interpretability and computational efficiency. On the negative side, however, using highly misspecified models can lead to unfairness, lack of validity, and poor downstream decisions. For example, modeling may be too coarse and not represent certain demographic groups properly. Going forward, it would be valuable to develop deeper understanding of such negative aspects of modeling. Overall, given that models are ubiquitous in practice, we believe they merit further study—especially under misspecification—and we have only scratched the surface of this agenda.

References

Brown, G., Hod, S., and Kalemaj, I. Performative prediction in a stateful world. In *International Conference on Artificial Intelligence and Statistics*, pp. 6045–6061. PMLR, 2022.

Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M., Zhang, K., and Zhao, L. Models as approximations I: Consequences illustrated with linear regression. *Statistical Science*, 34(4):523–544, 2019a.

Buja, A., Brown, L., Kuchibhotla, A. K., Berk, R., George, E., and Zhao, L. Models as approximations II: A model-free theory of parametric regression. *Statistical Science*, 34(4):545–565, 2019b.

Chen, Y., Liu, Y., and Podimata, C. Learning strategy-aware linear classifiers. *Advances in Neural Information Processing Systems*, 33:15265–15276, 2020.

Dean, S., Curmei, M., Ratliff, L. J., Morgenstern, J., and Fazel, M. Multi-learner risk reduction under endogenous participation dynamics. *arXiv preprint arXiv:2206.02667*, 2022.

Dong, J., Roth, A., Schutzman, Z., Waggoner, B., and Wu, Z. S. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 55–70, 2018.

Drusvyatskiy, D. and Xiao, L. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 2022.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. Online convex optimization in the bandit setting: gradient descent without a gradient. In *ACM-SIAM Symposium on Discrete Algorithms*, 2004.

Geer, S. A. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Ghalme, G., Nair, V., Eilat, I., Talgam-Cohen, I., and Rosenfeld, N. Strategic classification in the dark. In *International Conference on Machine Learning*, pp. 3672–3681. PMLR, 2021.

Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. Strategic classification. In *Proceedings of the 2016*

ACM conference on innovations in theoretical computer science, pp. 111–122, 2016.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Izzo, Z., Ying, L., and Zou, J. How to learn when data reacts to your model: performative gradient descent. In *International Conference on Machine Learning*, pp. 4641–4650. PMLR, 2021.

Izzo, Z., Zou, J., and Ying, L. How to learn when data gradually reacts to your model. In *International Conference on Artificial Intelligence and Statistics*, pp. 3998–4035. PMLR, 2022.

Jagadeesan, M., Mendler-Dünner, C., and Hardt, M. Alternative microfoundations for strategic classification. In *International Conference on Machine Learning*, pp. 4687–4697. PMLR, 2021.

Jagadeesan, M., Zrnic, T., and Mendler-Dünner, C. Regret minimization with performative feedback. In *International Conference on Machine Learning*, pp. 9760–9785. PMLR, 2022.

Kennedy, E. H. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

Kim, M. P. and Perdomo, J. C. Making decisions under outcome performativity. In *14th Innovations in Theoretical Computer Science Conference (ITCS)*, 2023.

Kleinberg, J. and Raghavan, M. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)*, 8(4):1–23, 2020.

Kleinberg, R., Slivkins, A., and Upfal, E. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pp. 681–690, 2008.

Levanon, S. and Rosenfeld, N. Strategic classification made practical. In *International Conference on Machine Learning*, pp. 6243–6253. PMLR, 2021.

Li, Q. and Wai, H.-T. State dependent performative prediction with stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3164–3186. PMLR, 2022.

Maheshwari, C., Chiu, C.-Y., Mazumdar, E., Sastry, S., and Ratliff, L. Zeroth-order methods for convex-concave minmax problems: Applications to decision-dependent risk minimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 6702–6734. PMLR, 2022.

Mendler-Dünner, C., Perdomo, J., Zrnic, T., and Hardt, M. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33: 4929–4939, 2020.

Mendler-Dünner, C., Ding, F., and Wang, Y. Anticipating performativity by predicting from predictions. *Advances in Neural Information Processing Systems*, 35:31171–31185, 2022.

Miller, J. P., Perdomo, J. C., and Zrnic, T. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, pp. 7710–7720. PMLR, 2021.

Mou, W., Ho, N., Wainwright, M. J., Bartlett, P., and Jordan, M. I. A diffusion process perspective on posterior contraction rates for parameters. *arXiv preprint arXiv:1909.00966*, 2019.

Narang, A., Faulkner, E., Drusvyatskiy, D., Fazel, M., and Ratliff, L. Learning in stochastic monotone games with decision-dependent data. In *International Conference on Artificial Intelligence and Statistics*, pp. 5891–5912. PMLR, 2022.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020.

Piliouras, G. and Yu, F.-Y. Multi-agent performative prediction: From global stability and optimality to chaos. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pp. 1047–1074, 2023.

Qiang, L., Yau, C.-Y., and Wai, H. T. Multi-agent performative prediction with greedy deployment and consensus seeking agents. In *Advances in Neural Information Processing Systems*, 2022.

Ray, M., Ratliff, L. J., Drusvyatskiy, D., and Fazel, M. Decision-dependent risk minimization in geometrically decaying dynamic environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8081–8088, 2022.

Spence, M. Job market signaling. In *Uncertainty in economics*, pp. 281–306. Elsevier, 1978.

Tsiatis, A. *Semiparametric Theory and Missing Data*. New York: Springer, 2007.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

White, H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48:817–838, 1980.

White, H. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912526.

Wooldridge, M. *Reasoning about rational agents*. MIT press, 2003.

Zrnic, T., Mazumdar, E., Sastry, S., and Jordan, M. Who leads and who follows in strategic classification? *Advances in Neural Information Processing Systems*, 34: 15257–15269, 2021.

A. Proofs

A.1. Notation and Definitions

In the proofs we will sometimes use $c, c' > 0$ to denote universal constants and $C, C' > 0$ to denote constants that may depend on the parameters introduced in the assumptions. We allow the values of the constants to vary from place to place.

We say a random variable x is *subexponential* with parameter ν if

$$\mathbb{P}\{|x| \geq t\} \leq 2\exp\left(-\frac{t}{\nu}\right)$$

for any $t \geq 0$. Unless specified, we do not assume x has mean zero in general. Moreover, we say a vector $\mathbf{x} \in \mathbb{R}^d$ is subexponential with parameter ν if $\langle u, \mathbf{x}\rangle$ is subexponential with parameter ν for any fixed direction $u \in \mathcal{S}^{d-1}$. Similarly, we say a random variable x is *subgaussian* with parameter σ if

$$\mathbb{P}\{|x| \geq t\} \leq 2\exp\left(-\frac{t^2}{\sigma^2}\right)$$

for any $t \geq 0$. Likewise, a vector $\mathbf{x} \in \mathbb{R}^d$ is subgaussian with parameter σ if $\langle u, \mathbf{x}\rangle$ is subgaussian with parameter σ for any fixed direction $u \in \mathcal{S}^{d-1}$.

A.2. Proof of Theorem 3.1

Define the population-level counterpart of $\hat{\theta}_{\mathrm{PO}}$ as:

$$\theta^* = \arg\min_{\theta \in \Theta} \mathrm{PR}^{\beta^*}(\theta).$$

We can write

$$\begin{aligned}
&\mathrm{PR}(\hat{\theta}_{\mathrm{PO}}) - \mathrm{PR}(\theta_{\mathrm{PO}}) \\
&\quad = (\mathrm{PR}(\hat{\theta}_{\mathrm{PO}}) - \mathrm{PR}^{\beta^*}(\hat{\theta}_{\mathrm{PO}})) + (\mathrm{PR}^{\beta^*}(\hat{\theta}_{\mathrm{PO}}) - \mathrm{PR}^{\hat{\beta}}(\hat{\theta}_{\mathrm{PO}})) + (\mathrm{PR}^{\hat{\beta}}(\hat{\theta}_{\mathrm{PO}}) - \mathrm{PR}^{\hat{\beta}}(\theta^*)) \\
&\quad + (\mathrm{PR}^{\hat{\beta}}(\theta^*) - \mathrm{PR}^{\beta^*}(\theta^*)) + (\mathrm{PR}^{\beta^*}(\theta^*) - \mathrm{PR}^{\beta^*}(\theta_{\mathrm{PO}})) + (\mathrm{PR}^{\beta^*}(\theta_{\mathrm{PO}}) - \mathrm{PR}(\theta_{\mathrm{PO}})).
\end{aligned}$$

By the definition of $\hat{\theta}_{\mathrm{PO}}$, we know $\mathrm{PR}^{\hat{\beta}}(\hat{\theta}) - \mathrm{PR}^{\hat{\beta}}(\theta^*) \leq 0$. Similarly, by the definition of θ^*, we know $\mathrm{PR}^{\beta^*}(\theta^*) - \mathrm{PR}^{\beta^*}(\theta_{\mathrm{PO}}) \leq 0$. Using these inequalities, we establish

$$\begin{aligned}
\mathrm{PR}(\hat{\theta}_{\mathrm{PO}}) - \mathrm{PR}(\theta_{\mathrm{PO}}) &\leq (\mathrm{PR}(\hat{\theta}_{\mathrm{PO}}) - \mathrm{PR}^{\beta^*}(\hat{\theta}_{\mathrm{PO}})) + (\mathrm{PR}^{\beta^*}(\hat{\theta}_{\mathrm{PO}}) - \mathrm{PR}^{\hat{\beta}}(\hat{\theta}_{\mathrm{PO}})) \\
&\quad + (\mathrm{PR}^{\hat{\beta}}(\theta^*) - \mathrm{PR}^{\beta^*}(\theta^*)) + (\mathrm{PR}^{\beta^*}(\theta_{\mathrm{PO}}) - \mathrm{PR}(\theta_{\mathrm{PO}})) \\
&\leq 2\sup_{\theta} |\mathrm{PR}(\theta) - \mathrm{PR}^{\beta^*}(\theta)| + 2\sup_{\theta} |\mathrm{PR}^{\beta^*}(\theta) - \mathrm{PR}^{\hat{\beta}}(\theta)| \\
&= 2(\mathtt{StatErr}_n + \mathtt{MisspecErr}).
\end{aligned}$$

A.3. Proof of Corollary 3.5

We have

$$\begin{aligned}
|\mathrm{PR}^{\beta^*}(\theta) - \mathrm{PR}(\theta)| &= \left|\int \ell(z;\theta)(p_{\beta^*}(z;\theta) - p(z;\theta))dz\right| \\
&\leq B_\ell \cdot \int |p_{\beta^*}(z;\theta) - p(z;\theta)|dz \\
&= 2B_\ell \cdot \mathrm{TV}(\mathcal{D}_{\beta^*}(\theta), \mathcal{D}(\theta)).
\end{aligned}$$

Therefore,

$$\mathtt{MisspecErr} \leq 2B_\ell \sup_{\theta \in \Theta} \mathrm{TV}(\mathcal{D}_{\beta^*}(\theta), \mathcal{D}(\theta)) \leq 2B_\ell \eta_{\mathrm{TV}}.$$

By a similar argument as above, $|\mathrm{PR}^{\beta^*}(\theta) - \mathrm{PR}^{\hat{\beta}}(\theta)| \leq 2B_\ell \mathrm{TV}(\mathcal{D}_{\beta^*}(\theta), \mathcal{D}_{\hat{\beta}}(\theta))$. Applying ϵ_{TV}-smoothness of the distribution atlas, we get

$$\texttt{StatErr}_n \leq 2B_\ell \sup_{\theta\in\Theta} \mathrm{TV}(\mathcal{D}_{\beta^*}(\theta), \mathcal{D}_{\hat{\beta}}(\theta)) \leq 2B_\ell \epsilon_{\mathrm{TV}} \|\hat{\beta} - \beta^*\| \leq 2B_\ell \epsilon_{\mathrm{TV}} C_n.$$

Applying Theorem 3.1 completes the proof.

A.4. Proof of Corollary 3.6

Denote by $\Pi(\mathcal{D}, \mathcal{D}')$ a coupling between two distributions $\mathcal{D}$ and $\mathcal{D}'$. We have

$$\begin{aligned}
\left|\mathrm{PR}^{\beta^*}(\theta) - \mathrm{PR}(\theta)\right| &= \left|\inf_{\Pi(\mathcal{D}_{\beta^*}(\theta), \mathcal{D}(\theta))} \mathbb{E}_{(z,z')\sim\Pi(\mathcal{D}_{\beta^*}(\theta), \mathcal{D}(\theta))}[\ell(z;\theta) - \ell(z';\theta)]\right| \\
&\leq \inf_{\Pi(\mathcal{D}_{\beta^*}(\theta), \mathcal{D}(\theta))} \mathbb{E}_{(z,z')\sim\Pi(\mathcal{D}_{\beta^*}(\theta), \mathcal{D}(\theta))}[|\ell(z;\theta) - \ell(z';\theta)|] \\
&\leq L \inf_{\Pi(\mathcal{D}_{\beta^*}(\theta), \mathcal{D}(\theta))} \mathbb{E}_{(z,z')\sim\Pi(\mathcal{D}_{\beta^*}(\theta), \mathcal{D}(\theta))}[\|z - z'\|] \\
&= L\mathcal{W}(\mathcal{D}_{\beta^*}(\theta), \mathcal{D}(\theta)).
\end{aligned}$$

Therefore,

$$\texttt{MisspecErr} \leq L \sup_{\theta\in\Theta} \mathcal{W}(\mathcal{D}_{\beta^*}(\theta), \mathcal{D}(\theta)) \leq L \cdot \eta_W.$$

By a similar argument, $|\mathrm{PR}^{\beta^*}(\theta) - \mathrm{PR}^{\hat{\beta}}(\theta)| \leq L\mathcal{W}(\mathcal{D}_{\beta^*}(\theta), \mathcal{D}_{\hat{\beta}}(\theta))$. Applying ϵ_W-smoothness of the distribution atlas, we get

$$\texttt{StatErr}_n \leq L \sup_{\theta\in\Theta} \mathcal{W}(\mathcal{D}_{\beta^*}(\theta), \mathcal{D}_{\hat{\beta}}(\theta)) \leq L\epsilon_W \|\hat{\beta} - \beta^*\| \leq L \cdot \epsilon_W \cdot C_n.$$

Applying Theorem 3.1 completes the proof.

A.5. Proof of Lemma 3.8

We first present Lemma A.1 which we will use in the proof.

Lemma A.1.

$$\langle \nabla r(\beta),\, \beta - \beta^* \rangle \geq \begin{cases} \dfrac{\mu}{2}\|\beta^* - \beta\|^2, & \|\beta^* - \beta\| \leq \dfrac{\mu}{\sigma_r}, \\ \dfrac{\mu^2}{2\sigma_r}\|\beta^* - \beta\|, & \|\beta^* - \beta\| \geq \dfrac{\mu}{\sigma_r}. \end{cases} \tag{3}$$

See the proof of Lemma A.1 in Supplement A.6

Let B_β be the bound of the parameter set $\mathcal{B}$, i.e., $\|\beta\| \leq \mathcal{B}_\beta$ for any $\beta \in \mathcal{B}$. We will show Lemma 3.8 holds with some sufficiently large constants $C, C' > 0$ that depend polynomially on $(\log|1 + B_\beta|, 1/\mu, L_r, B_r, \sigma_r)$.

Denote

$$r_n(\beta) := \frac{1}{n}\sum_{i=1}^n r(\theta_i, z_i; \beta); \quad r(\beta) := \mathbb{E}_{\theta\sim\tilde{\mathcal{D}}, z\sim\mathcal{D}(\theta)}[r(\theta, z; \beta)].$$

We begin by claiming the following result, which we will prove later. With probability over $1 - \delta$

$$\sup_{\beta\in\mathcal{B}} \|\nabla_\beta r_n(\beta) - \nabla_\beta r(\beta)\| \leq C\sqrt{\log n}\sqrt{\frac{d_\beta + \log(1/\delta)}{n}}. \tag{4}$$

With this result at hand, it follows from Lemma A.1 that

$$\begin{aligned}
\min\left\{\frac{\mu}{2}\|\hat{\beta} - \beta^*\|^2, \frac{\mu^2}{2\sigma_r}\|\hat{\beta} - \beta^*\|\right\} &\leq \langle \nabla_\beta r(\hat{\beta}),\, \hat{\beta} - \beta^* \rangle = \langle \nabla_\beta r(\hat{\beta}) - \nabla_\beta r_n(\hat{\beta}),\, \hat{\beta} - \beta^* \rangle \\
&\leq \|\nabla_\beta r(\hat{\beta}) - \nabla_\beta r_n(\hat{\beta})\|\|\hat{\beta} - \beta^*\| \leq C\sqrt{\log n}\sqrt{\frac{d_\beta + \log(1/\delta)}{n}}\|\hat{\beta} - \beta^*\|,
\end{aligned} \tag{5}$$

where the first equality is due to the fact that $\nabla_\beta r_n(\hat{\beta}) = 0$. Eliminating $\|\hat{\beta} - \beta^*\|$ in both the first and the last term of (5) yields

$$\frac{\mu^2}{2\sigma_r} \wedge \frac{\mu}{2}\|\hat{\beta} - \beta^*\| \le C\sqrt{\log n}\sqrt{\frac{d_\beta + \log(1/\delta)}{n}}.$$

By the sample size Assumption in Lemma 3.8, we may assume n is sufficiently large such that $\frac{\mu^2}{2\sigma_r} \ge C\sqrt{\log n}\sqrt{\frac{d_\beta+\log(1/\delta)}{n}}$ for some constant $C > 0$. Therefore, we conclude that

$$\|\hat{\beta} - \beta^*\| \le C\sqrt{\log n}\sqrt{\frac{d_\beta + \log(1/\delta)}{n}}.$$

Proof of Eq. (4). Let $\{u_1, u_2, \dots, u_M\}$ be a $1/2$-covering of $\mathcal{S}^{d_\beta - 1}$ in the Euclidean norm such that $|M| \le 5^{d_\beta}$. Define the random variables

$$\phi_{u,\beta} := \langle u,\, \nabla_\beta r_n(\beta) - \nabla_\beta r(\beta)\rangle, \;\; \phi_u := \sup_{\beta\in\mathcal{B}} \phi_{u,\beta}.$$

It follows from a standard discretization argument (e.g., (Wainwright, 2019), Chap. 6) that

$$\sup_{\beta\in\mathcal{B}} \|\nabla_\beta r_n(\beta) - \nabla_\beta r(\beta)\| \le 2\sup_{\beta\in\mathcal{B}}\sup_{i\in[M]} \phi_{u_i,\beta}. \tag{6}$$

We make the following claim which will be proved at the end. With probability over $1-\delta$

$$\left\|\frac{1}{n}\sum_{i=1}^n \sup_{\beta\in\mathcal{B}} \nabla^2_\beta r(\theta_i, z_i; \beta)\right\|_{\text{op}} \le cL_r + cL_r\sqrt{\frac{d_\beta + \log(1/\delta)}{n}} \le cL_r. \tag{7}$$

for some constant $c > 0$.

Let $\epsilon > 0$ be some value we specify later. Construct an ε-covering net $\{\beta^1, \dots, \beta^N\}$ of $\mathcal{B}$ in $\|\cdot\|$. Then the covering number $|N| \le (1 + 2B_\beta/\epsilon)^{d_\beta}$, and

$$\begin{aligned}
\sup_{\beta\in\mathcal{B}}\sup_{i\in[M]} \phi_{u_i,\beta} &\le \sup_{i\in[M]}\sup_{j\in[N]} \phi_{u_i,\beta^j} + \sup_{i\in[M]}\sup_{\|\beta_1-\beta_2\|\le\varepsilon} |\phi_{u_i,\beta_1} - \phi_{u_i,\beta_2}| \\
&\le \sup_{i\in[M]}\sup_{j\in[N]} \phi_{u_i,\beta^j} + \sup_{i\in[M]}\sup_{\|\beta_1-\beta_2\|\le\varepsilon} \frac{1}{n}\sum_{i=1}^n \sup_{\beta\in\mathcal{B}} u_i^\top \nabla^2_\beta r(\theta_i, z_i; \beta)(\beta_1 - \beta_2) \\
&\le \sup_{i\in[M]}\sup_{j\in[N]} \phi_{u_i,\beta^j} + \sup_{i\in[M]}\sup_{\|\beta_1-\beta_2\|\le\varepsilon} \|u_i\| \left\|\frac{1}{n}\sum_{i=1}^n \sup_{\beta\in\mathcal{B}} \nabla^2_\beta r(\theta_i, z_i; \beta)\right\|_{\text{op}} \|\beta_1 - \beta_2\| \\
&\le \sup_{i\in[M]}\sup_{j\in[N]} \phi_{u_i,\beta^j} + cL_r\varepsilon,
\end{aligned} \tag{8}$$

where the last inequality follows from the claim in (7). Since $\langle u,\, \nabla_\beta r(\theta_i, z_i; \beta) - \mathbb{E}[\nabla_\beta r(\theta_i, z_i; \beta)]\rangle$ is zero mean subexponential with parameter B_r by condition (b) of Assumption 3.7, it follows from concentration of subexponential variables that

$$\mathbb{P}\{|\phi_{u_i,\beta^j}| \ge t\} \le 2\exp\left(-\frac{nt^2}{2B_r^2}\right) \quad \text{for any } |t| \le B_r.$$

Applying a union bound over i, j, we establish

$$\mathbb{P}\left\{\max_{i\in[M], j\in[N]} |\phi_{u_i,\beta^j}| \ge t\right\} \le 2\exp\left(cd_\beta(\log(1 + 2B_\beta/\epsilon) + 1) - \frac{nt^2}{2B_r^2}\right) \quad \text{for any } |t| \le B_r. \tag{9}$$

Let $\epsilon = \sqrt{\frac{d_\beta+\log(1/\delta)}{n}}$ and

$$t = \frac{cB_r\sqrt{\log(1/\delta) + d_\beta(\log(1 + 2B_\beta/\epsilon) + 1)}}{\sqrt{n}} \le \frac{CB_r\sqrt{\log n(d_\beta + \log(1/\delta))}}{\sqrt{n}} \le B_r$$

for some constant $c > 0$, where the last inequality uses the sample size assumption of the lemma. Substituting the values of ϵ and t into Equations (8), (9) and combining with Eq. (6), we obtain

$$\begin{aligned}\sup_{\beta\in\mathcal{B}}\|\nabla_\beta r_n(\beta)-\nabla_\beta r(\beta)\| &\le \frac{CB_r\sqrt{\log n(d_\beta+\log(1/\delta))}}{\sqrt{n}}+cL_r\sqrt{\frac{d_\beta+\log(1/\delta)}{n}}\\ &\le C'\sqrt{\log n}\sqrt{\frac{d_\beta+\log(1/\delta)}{n}}\end{aligned}$$

with probability over $1-\delta$ for some parameter-dependent constant C'.

Proof of Eq. (7). Similar to equation (6), from a standard discretization argument we have

$$\left\|\frac{1}{n}\sum_{i=1}^n\sup_{\beta\in\mathcal{B}}\nabla^2_\beta r(\theta_i,z_i;\beta)\right\|_{\mathrm{op}}\le 2\sup_{j\in[M]}\frac{1}{n}\sum_{i=1}^n\sup_{\beta\in\mathcal{B}}u_j^\top\nabla^2_\beta r(\theta_i,z_i;\beta)u_j.$$

Since $\sup_{\beta\in\mathcal{B}}u_j^\top\nabla^2_\beta r(\theta_i,z_i;\beta)u_j$ are subexponential variables by condition (c) in Assumption 3.7, we have from properties of subexponential variables and Bernstein's inequality that

$$\begin{aligned}&\mathbb{P}\left\{u_j^\top\frac{1}{n}\sum_{i=1}^n\sup_{\beta\in\mathcal{B}}\nabla^2_\beta r(\theta_i,z_i;\beta)u_j\ge cL_r+t\right\}\\ &\le\mathbb{P}\left\{u_j^\top\frac{1}{n}\sum_{i=1}^n\sup_{\beta\in\mathcal{B}}\nabla^2_\beta r(\theta_i,z_i;\beta)u_j\ge\mathbb{E}[u_j^\top\sup_{\beta\in\mathcal{B}}\nabla^2_\beta r(\theta_i,z_i;\beta)u_j]+t\right\}\\ &\le\exp\left(-c\min\{\frac{nt}{L_r},\frac{nt^2}{L_r^2}\}\right).\end{aligned}$$

Applying a union bound over $j\in[M]$ and setting $t=cL_r\sqrt{\frac{d_\beta+\log(1/\delta)}{n}}<cL_r$, we establish

$$u_j^\top\frac{1}{n}\sum_{i=1}^n\sup_{\beta\in\mathcal{B}}\nabla^2_\beta r(\theta_i,z_i;\beta)u_j\le cL_r+cL_r\sqrt{\frac{d_\beta+\log(1/\delta)}{n}}\le cL_r$$

for some $c>0$ with probability over $1-\delta$.

A.6. Proof of Lemma A.1

For any $\|\beta^*-\beta\|\le\frac{\mu}{\sigma_r}$, by a Taylor expansion of $\nabla r(\beta)$ at β^* and Ass. 3.7(a), we have

$$\langle\nabla r(\beta),\,\beta-\beta^*\rangle=\langle\nabla r(\beta)-\nabla r(\beta^*),\,\beta-\beta^*\rangle\ge\mu\|\beta-\beta^*\|^2-\frac{\sigma_r}{2}\|\beta^*-\beta\|^3\ge\frac{\mu}{2}\|\beta^*-\beta\|^2.$$

This gives the second part of Lemma A.1. When $\|\beta^*-\beta\|\ge\frac{\mu}{\sigma_r}$, write $\beta=\beta^*+tu$, where $u=(\beta-\beta^*)/\|\beta-\beta^*\|$. For a fixed direction u, define $\beta(t):=\beta^*+tu$ and

$$f(u,t):=\left\langle\nabla r(\beta(t)),\frac{\beta(t)-\beta^*}{\|\beta(t)-\beta^*\|}\right\rangle=\langle u,\,\nabla r(\beta(t))\rangle.$$

Then for $t\ge0$, using Ass. 3.7(a) again we obtain

$$\partial_t f(u,t)=\langle u,\,\nabla^2 r(\beta(t))u\rangle\ge0$$

Therefore $f(u,t)$ is increasing in t and for any $\|\beta-\beta^*\|\ge\frac{\mu}{\sigma_r}$

$$\begin{aligned}\left\langle\nabla r(\beta),\frac{\beta-\beta^*}{\|\beta-\beta^*\|}\right\rangle&\ge\left\langle\nabla r\Big(\beta\Big(\frac{\mu}{\sigma_r}\Big)\Big),\frac{\beta\Big(\frac{\mu}{\sigma_r}\Big)-\beta^*}{\|\beta\Big(\frac{\mu}{\sigma_r}\Big)-\beta^*\|}\right\rangle\\ &\ge\frac{\mu}{2}\|\beta\Big(\frac{\mu}{\sigma_r}\Big)-\beta^*\|=\frac{\mu^2}{2\sigma_r}.\end{aligned}$$

This gives the second part of Lemma A.1.

A.7. Proof of Theorem 3.9

The first statement follows directly by putting together Corollary 3.5 and Lemma 3.8. Similarly, the second statement follows by putting together Corollary 3.6 and Lemma 3.8.

A.8. Proof of Claim 4.1

Fix x and θ. We will use the shorthand notation $g_\beta(x,\theta) \equiv x_\beta$. First, we show that $\|x_\beta - x_{\beta'}\|_2 \leq \frac{B_u}{1-\beta_{\max} L_u}|\beta - \beta'|$. To see this, notice that the optimality condition of the best-response equation is equal to:

$$\beta \cdot \nabla u_\theta(x_\beta) - x_\beta = -x.$$

Since $\beta \nabla u_\theta(x_\beta) - x_\beta = \beta' \nabla u_\theta(x_{\beta'}) - x_{\beta'} = -x$, we know

$$\|\beta \nabla u_\theta(x_\beta) - \beta' \nabla u_\theta(x_{\beta'})\| = \|x_\beta - x_{\beta'}\|.$$

We also know

$$\begin{aligned}\|\beta \nabla u_\theta(x_\beta) - \beta' \nabla u_\theta(x_{\beta'})\| &= \|\beta \nabla u_\theta(x_\beta) - \beta \nabla u_\theta(x_{\beta'}) + \beta \nabla u_\theta(x_{\beta'}) - \beta' \nabla u_\theta(x_{\beta'})\| \\ &\leq \beta L_u \|x_\beta - x_{\beta'}\| + B_u \left|\beta - \beta'\right|.\end{aligned}$$

Therefore,

$$\|x_\beta - x_{\beta'}\| \leq \beta L_u \|x_\beta - x_{\beta'}\| + B_u \left|\beta - \beta'\right|.$$

Rearranging the terms, we get

$$\|x_\beta - x_{\beta'}\| \leq \frac{B_u}{1 - \beta L_u}|\beta - \beta'|.$$

By the definition of Wasserstein distance, this condition directly implies

$$\mathcal{W}(\mathcal{D}_\beta(\theta), \mathcal{D}_{\beta'}(\theta)) \leq \frac{B_u}{1 - \beta_{\max} L_u}|\beta - \beta'|,$$

which is the definition of $\frac{B_u}{1-\beta_{\max} L_u}$-smoothness.

A.9. Proof of Claim 4.2

The claim follows by Lemma 3.8 after verifying the conditions required in Assumption 3.7. We have $r(\theta, x; \beta) = \|x - \beta \nabla u_\theta(x)\|^2$, so $\nabla r(\theta, x; \beta) = -2(x^\top \nabla u_\theta(x) - \beta \|\nabla u_\theta(x)\|^2)$ and $\nabla^2 r(\theta, x; \beta) = 2\|\nabla u_\theta(x)\|^2$. Conditions (b) and (c) of Assumption 3.7 are thus satisfied by $\tilde{x}$ and $\nabla u_{\tilde{\theta}}(\tilde{x})$ being subgaussian since products of subgaussians are subexponential. Condition (a) is satisfied by the fact that $r(\beta) = \mathbb{E}[\|\tilde{x} - \beta \nabla u_{\tilde{\theta}}(\tilde{x})\|^2]$ is a quadratic in β when $\mathbb{E}[\|\nabla u_{\tilde{\theta}}(\tilde{x})\|^2] > 0$.

A.10. Proof of Claim 4.3

Fix θ, β, β', and without loss of generality let $\beta > \beta'$. We show that $\mathrm{TV}(\mathcal{D}_\beta(\theta), \mathcal{D}_{\beta'}(\theta)) \leq \phi_u$. The distributions $\mathcal{D}_\beta(\theta)$ and $\mathcal{D}_{\beta'}(\theta)$ are equal to each other and to $\mathcal{D}_0$ for all $\{x : x^\top \theta \in (-\infty, T - \beta) \cup (T, \infty)\}$. Moreover, under both $\mathcal{D}_\beta(\theta)$ and $\mathcal{D}_{\beta'}(\theta)$, there is no mass for $\{x : x^\top \theta \in (T - \beta', T)\}$. The distributions thus only differ for $\{x : x^\top \theta \in [T - \beta, T - \beta'] \cup \{T\}\}$. Since the density of $x^\top \theta$ is bounded by ϕ_u, the measure of such vectors x is at most $\phi_u |\beta - \beta'|$.

A.11. Proof of Claim 4.4

By Hoeffding's inequality, with probability $1 - \delta$ it holds that

$$\left| \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{x_i^\top \theta_i \in (T \pm \epsilon)\} - \mathbb{P}\{\tilde{x}^\top \tilde{\theta} \in (T \pm \epsilon)\} \right| \leq \sqrt{\frac{\log(2/\delta)}{2n}}.$$

Let $\Delta = \sqrt{\frac{\log(2/\delta)}{2n}}$. Next we argue that $|\hat{\beta} - \beta^*| \leq \frac{\Delta}{\phi_l}$ by contradiction. Suppose $|\hat{\beta} - \beta^*| > \frac{\Delta}{\phi_l}$. Then,

$$\begin{aligned}
\left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{x_i^\top \theta_i \in (T \pm \epsilon)\} - \mathbb{P}\{\tilde{x}^\top \tilde{\theta} \in (T \pm \epsilon)\} \right| &= \left| \mathbb{P}\{x_0^\top \tilde{\theta} \in [T - \hat{\beta}, T]\} - \mathbb{P}\{x_0^\top \tilde{\theta} \in [T - \beta^*, T]\} \right| \\
&> \phi_l \frac{\Delta}{\phi_l} \\
&= \Delta,
\end{aligned}$$

which contradicts Hoeffding's inequality. Therefore, we conclude that $|\hat{\beta} - \beta^*| \leq \frac{\Delta}{\phi_l}$.

A.12. Proof of Claim 4.5

By the definition of Wasserstein distance, we have

$$\mathcal{W}(\mathcal{D}_{\mathcal{M}_1}(\theta), \mathcal{D}_{\mathcal{M}_2}(\theta)) = \mathcal{W}(\mathcal{M}_1\theta + z_0, \mathcal{M}_2\theta + z_0) = \|\mathcal{M}_1\theta - \mathcal{M}_2\theta\| \leq B_\theta \left\|\mathcal{M}_1 - \mathcal{M}_2\right\|_{\text{op}}$$

for any $\mathcal{M}_1, \mathcal{M}_2$. Therefore, the distribution atlas $\{\mathcal{D}_\mathcal{M}\}_\mathcal{M}$ is ϵ_W-smooth with parameter B_θ.

A.13. Proof of Claim 4.6

Let ν_θ be the subgaussian parameter of $\tilde{\mathcal{D}}$. We prove that there exists C, C' depending polynomially on $(1/\kappa_{\min}, \kappa_{\max}, \nu_\theta, L_{\theta z}, B)$ such that Claim 4.6 holds. By definition, we have

$$\begin{aligned}
\widehat{\mathcal{M}}^\top &= \left(\frac{1}{n}\sum_{i=1}^{n} \theta_i \theta_i^\top\right)^{-1} \Big(\frac{1}{n}\sum_{i=1}^{n} \theta_i z_i^\top\Big), \\
\mathcal{M}^{*\top} &= \mathbb{E}[\tilde{\theta}\tilde{\theta}^\top]^{-1} \mathbb{E}[\tilde{\theta}\tilde{z}^\top].
\end{aligned}$$

We state the following results, which we will prove later:

$$\left\| \left(\frac{1}{n}\sum_{i=1}^{n} \theta_i \theta_i^\top\right)^{-1} - \mathbb{E}[\theta_i\theta_i^\top]^{-1} \right\|_{\text{op}} \leq \frac{c\nu_\theta^2}{\kappa_{\min}^2} \sqrt{\frac{d_\theta + \log(1/\delta)}{n}} \leq C; \tag{10}$$

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \theta_i z_i^\top - \mathbb{E}[\theta_i z_i^\top] \right\|_{\text{op}} \leq cL_{\theta,z} \sqrt{\frac{d_\theta + d_z + \log(1/\delta)}{n}}. \tag{11}$$

Combining Equations (10), (11) with the assumptions of the claim, we establish

$$\begin{aligned}
\left\|\widehat{\mathcal{M}} - \mathcal{M}^*\right\|_{\text{op}} &= \left\| \left(\frac{1}{n}\sum_{i=1}^{n} \theta_i \theta_i^\top\right)^{-1} \Big(\frac{1}{n}\sum_{i=1}^{n} \theta_i z_i^\top\Big) - \mathbb{E}[\theta_i\theta_i^\top]^{-1}\mathbb{E}[\theta_i z_i^\top] \right\|_{\text{op}} \\
&\leq \left\| \left(\frac{1}{n}\sum_{i=1}^{n} \theta_i \theta_i^\top\right)^{-1} - \mathbb{E}[\theta_i\theta_i^\top]^{-1} \right\|_{\text{op}} \left\| \frac{1}{n}\sum_{i=1}^{n} \theta_i z_i^\top \right\|_{\text{op}} \\
&\quad + \left\|\mathbb{E}[\theta_i\theta_i^\top]^{-1}\right\|_{\text{op}} \left\| \frac{1}{n}\sum_{i=1}^{n} \theta_i z_i^\top - \mathbb{E}[\theta_i z_i^\top] \right\|_{\text{op}} \\
&\leq C' \sqrt{\frac{d_\theta + d_z + \log(1/\delta)}{n}}.
\end{aligned}$$

for some $C' > 0$ that depends on problem-specific parameters.

Proof of Eq. (10). Under the conditions of the claim, we establish from concentration inequalities for subgaussian vectors (see, e.g., Theorem 6.5 in (Wainwright, 2019)) that with probability at least $1-\delta$,

$$\begin{aligned}\frac{\kappa_{\min}}{2} \le \kappa_{\min} - c\nu_\theta^2\sqrt{\frac{d_\theta + \log(1/\delta)}{n}} &\le \sigma_{\min}\left(\frac{1}{n}\sum_{i=1}^n \theta_i\theta_i^\top\right), \\ &\le \sigma_{\max}\left(\frac{1}{n}\sum_{i=1}^n \theta_i\theta_i^\top\right) \le \kappa_{\max} + c\nu_\theta^2\sqrt{\frac{d_\theta + \log(1/\delta)}{n}} \le \frac{3}{2}\kappa_{\max},\end{aligned}$$

where the last line follows from the sample-size assumption. In addition, we also have from (Wainwright, 2019) that

$$\left\|\frac{1}{n}\sum_{i=1}^n \theta_i\theta_i^\top - \mathbb{E}[\theta_i\theta_i^\top]\right\|_{\text{op}} \le c\nu_\theta^2\sqrt{\frac{d_\theta + \log(1/\delta)}{n}}.$$

Therefore, it follows from Woodbury's matrix identity and the last two displays that

$$\begin{aligned}\left\|\left(\frac{1}{n}\sum_{i=1}^n \theta_i\theta_i^\top\right)^{-1} - \mathbb{E}[\theta_i\theta_i^\top]^{-1}\right\|_{\text{op}} &= \left\|\left(\frac{1}{n}\sum_{i=1}^n \theta_i\theta_i^\top\right)^{-1}\left(\left(\frac{1}{n}\sum_{i=1}^n \theta_i\theta_i^\top\right) - \mathbb{E}[\theta_i\theta_i^\top]\right)(\mathbb{E}[\theta_i\theta_i^\top])^{-1}\right\|_{\text{op}} \\ &\le \left\|\left(\frac{1}{n}\sum_{i=1}^n \theta_i\theta_i^\top\right)^{-1}\right\|_{\text{op}}\left\|\frac{1}{n}\sum_{i=1}^n \theta_i\theta_i^\top - \mathbb{E}[\theta_i\theta_i^\top]\right\|_{\text{op}}\left\|(\mathbb{E}[\theta_i\theta_i^\top])^{-1}\right\|_{\text{op}} \\ &\le \frac{c\nu_\theta^2}{\kappa_{\min}^2}\sqrt{\frac{d_\theta + \log(1/\delta)}{n}}.\end{aligned}$$

The second inequality follows from the assumption on sample size.

Proof of Eq. (11). Let $\{u_1, \ldots, u_M\}$ be a $1/4$-covering of $\mathcal{S}^{d_\theta - 1}$ in the Euclidean norm with $|M| \le 9^{d_\theta}$, and $\{v_1, \ldots, v_N\}$ to be a $1/4$-covering of $\mathcal{S}^{d_z - 1}$ with $|N| \le 9^{d_z}$. Then by a standard discretization argument, we have

$$\left\|\frac{1}{n}\sum_{i=1}^n \theta_i z_i^\top - \mathbb{E}[\theta_i z_i]\right\|_{\text{op}} \le 2\sup_{k\in[M], l\in[N]} \frac{1}{n}\sum_{i=1}^n u_k^\top\theta_i z_i^\top v_l - \mathbb{E}[u_k^\top\theta_i z_i^\top v_l].$$

Since $u_k^\top\theta_i z_i^\top v_l - \mathbb{E}[u_k^\top\theta_i z_i^\top v_l]$ are zero-mean subexponential variables by assumption, it follows that

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n u_k^\top\theta_i z_i^\top v_l - \mathbb{E}[u_k^\top\theta_i z_i^\top v_l]\right| \ge t\right\} \le 2\exp\left(-c\min\left\{\frac{nt^2}{L_{\theta z}^2}, \frac{nt}{L_{\theta z}}\right\}\right).$$

Applying a union bound over M, N and setting $t = cL_{\theta z}\sqrt{\frac{d_\theta + d_z + \log(1/\delta)}{n}} < cL_{\theta z}$ with some sufficiently large constant $c > 0$ yields

$$\begin{aligned}&\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^n u_k^\top\theta_i z_i^\top v_l - \mathbb{E}[u_k^\top\theta_i z_i^\top v_l]\right\|_{\text{op}} \ge t\right\} \\ &\le \mathbb{P}\left\{\sup_{k,l}\left|\frac{1}{n}\sum_{i=1}^n u_k^\top\theta_i z_i^\top v_l - \mathbb{E}[u_k^\top\theta_i z_i^\top v_l]\right| \ge t\right\} \\ &\le 2\exp\left((d_\theta + d_z)\log 9 - c\min\left\{\frac{nt^2}{L_{\theta z}^2}, \frac{nt}{L_{\theta z}}\right\}\right) \le \delta,\end{aligned}$$

which gives Eq. (11).

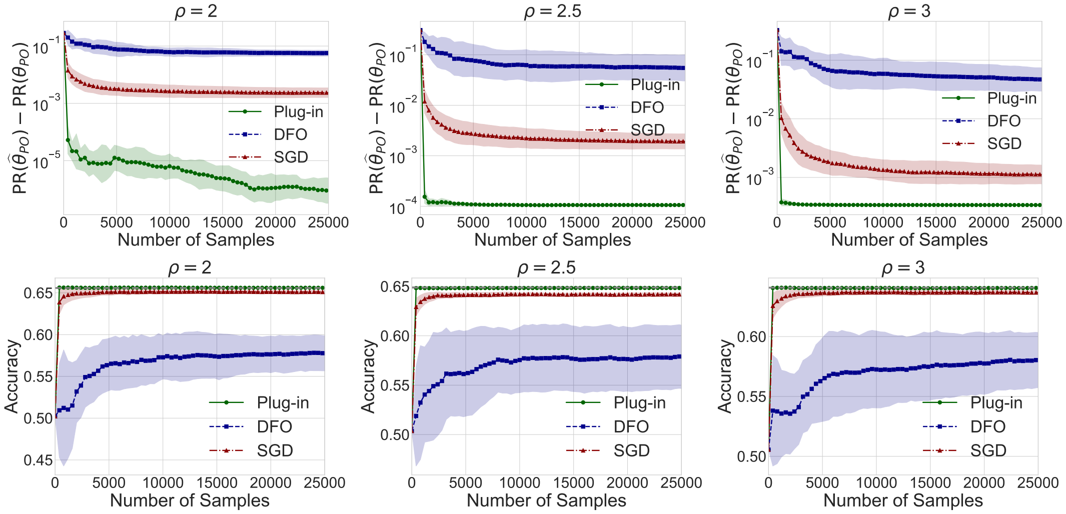

Figure 3. Excess risk (top) and accuracy (bottom) versus n for plug-in performative optimization, the DFO algorithm, and greedy SGD, with a changed value of $\tilde{\beta} = 1$. We display the ±1 standard deviation, logarithmically scaled. The takeaways are largely the same as in Figure 2.

B. Further Experimental Results and Details

We repeat each experiment 10 times and plot the mean excess risk as well as the ± 1 standard deviation. In all experiments on strategic classification, we choose the ridge parameter $\lambda = 0.001$.

In Figure 3 we provide an additional comparison in the context of the strategic-regression example from Section 5. We let $\tilde{\beta} = 1$, showing that our takeaways are robust to the exact value of $\tilde{\beta}$.

We also run the strategic-regression experiment on a real data set. We use the `credit` data set, in particular the processed version available at: https://github.com/ustunb/actionable-recourse. The data set contains $30,000$ samples of $d = 17$ features and a $\{0,1\}$-valued outcome y_i with $y_i = 1$ denoting individual i not defaulting on a credit card payment. The features include marital status, age, education level, and payment patterns. We assume the individuals can modify their records on education level and payment patterns (features 7–17), but cannot change other records. We use 1500 randomly drawn data points to form the base distribution $\mathcal{D}_0$; we assume the same true response model and use the same distribution atlas as before. We set $\tilde{\beta} = 5$, $\Theta = \{\theta : \|\theta\| \leq 1\}$, and standardize the features so that each column is zero-mean and has unit variance. In Figure 4, we observe patterns similar to those in Figure 2, though the gap in accuracy between our method and SGD is smaller.

Below we provide implementation details for the two considered baselines.

Derivative-free optimization (DFO). Starting from $\theta_0 = \mathbf{1}_d/\sqrt{d}$, we run the updates

$$\theta_{t+1} = \mathrm{Proj}_{\|\cdot\|\leq 1}(\theta_t - \eta_t \widehat{\mathbb{E}}[u \mathrm{PR}(\theta_t + \delta u) d/\delta])$$

for $t \geq 0$, where the step size $\eta_t = c_0/(t+1)$, u is uniformly distributed on $\mathcal{S}^{d-1}$, and $\widehat{\mathbb{E}}[u\mathrm{PR}(\theta_t + \delta u)d/\delta]$ denotes the unbiased sample estimation of $\mathbb{E}[u\mathrm{PR}(\theta_t + \delta u)d/\delta]$ using m i.i.d. pairs of $(u, z) \sim \mathrm{Unif}(\mathcal{S}^{d-1}) \times \mathcal{D}(\theta_t + \delta u)$. The projection $\mathrm{Proj}_{\|\cdot\|\leq 1}(x)$ denotes the projection of $x \in \mathbb{R}^d$ onto the ball $\{v \in \mathbb{R}^d : \|v\| \leq 1\}$ in Euclidean norm. We choose the step size parameter $c_0 \in [10^{-4}, 10^{-1}]$, the batch size m in $[1, 500]$, and $\delta \in [0.1, 100]$ via grid search.

Greedy stochastic gradient descent (SGD). Starting from $\theta_0 = 1_d/\sqrt{d}$, we run the updates

$$\theta_{t+1} = \mathrm{Proj}_{\|\cdot\|\leq 1}(\theta_t - \eta_t \nabla_\theta \ell(z_t; \theta_t))$$

with step size $\eta_t = c_0/(t+1)$ and $z_t \sim \mathcal{D}(\theta_t)$. The step size parameter $c_0 \in [10^{-4}, 10]$ and the batch size $m \in [1, 500]$ are selected via grid search. The greedy SGD algorithm neglects the implicit dependence of z on θ due to performativity, and therefore typically converges to suboptimal points.

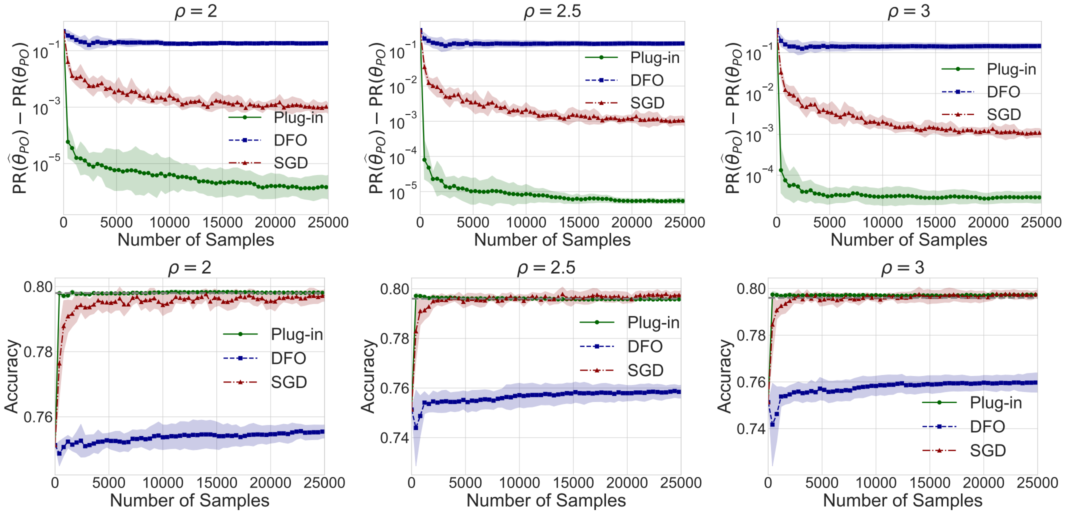

Figure 4. Excess risk (top) and accuracy (bottom) versus n for plug-in performative optimization, the DFO algorithm, and greedy SGD, on the `credit` data set. We display the ±1 standard deviation, logarithmically scaled.

Performative gradient descent (PerfGD). Assume the distribution map has the form $z \sim \mathcal{D}(\theta) \Leftrightarrow z \overset{d}{=} \mathcal{N}(f(\theta), \sigma^2 I_d)$, where f is some unknown smooth function. Starting from $\theta_0 = 1_d/\sqrt{d}$, we first run the greedy SGD updates for H burn-in steps. Next, we run SGD on the performative risk using an estimated performative gradient, namely,

$$\theta_{t+1} = \mathrm{Proj}_{\|\cdot\| \leq 1}(\theta_t - \eta_t \widehat{\nabla}_\theta \mathbb{E}[\ell(z_t; \theta_t)]),$$

with step size $\eta_t = c_0/(t+1)$ and $z_t \sim \mathcal{D}(\theta_t)$, where the estimated performative gradient is computed as in Algorithm 3 and Eq. (2) in (Izzo et al., 2021) via numerically estimating the gradient $\frac{\partial f}{\partial \theta}$.

We choose the number of burn-in steps $\mathrm{H} = 10d$. The step size parameter $c_0 \in [10^{-4}, 10]$ and the batch size $m \in [1, 500]$ are selected via grid search. PerfGD runs stochastic gradient descent on the performative risk using an estimated performative gradient. It should be noted that the numerical approximation of $\frac{\partial f}{\partial \theta}$ is unstable when $d > 1$, which results in the suboptimal performance of PerfGD in our location-family experiment.

C. Solving for $\hat{\theta}_{\mathrm{PO}}$

The map $\mathcal{D}_{\hat{\beta}}$ belongs to the distribution atlas chosen by the learner, and as such, it is fully specified and known to the learner. Therefore, solving for $\hat{\theta}_{\mathrm{PO}} = \arg\min_{\theta \in \Theta} \mathbb{E}_{z \sim \mathcal{D}_{\hat{\beta}}(\theta)}[\ell(z; \theta)]$ can only incur error due to computational inaccuracies. There is no additional statistical complexity (i.e. dependence on n), which is the focus of our excess risk bounds in Theorem 3.1. In a sense, our results can be thought of as analogous to classical generalization bounds for empirical risk minimizers: we are concerned with characterizing the performance of the empirical risk minimizer, not with computational strategies for finding them.

More practically, there are several approaches one can take to compute $\hat{\theta}_{\mathrm{PO}}$. Sometimes $\hat{\theta}_{\mathrm{PO}}$ has a closed-form expression, as in Example 1. In such cases there is no error in Step 3 of Algorithm 1. Sometimes $\mathcal{D}_{\hat{\beta}}(\theta)$ and $\ell(z; \theta)$ are simple enough that $\mathrm{PR}^{\hat{\beta}}(\theta)$ has a closed-form expression; in such cases, we compute $\hat{\theta}_{\mathrm{PO}}$ by running gradient descent on $\mathrm{PR}^{\hat{\beta}}(\theta)$. This is the case in all our experiments. Alternatively, if $\mathrm{PR}^{\hat{\beta}}(\theta)$ does not have a closed-form expression, one may compute $\hat{\theta}_{\mathrm{PO}}$ by using a black-box optimizer on an unbiased estimate of $\mathbb{E}_{z \sim \mathcal{D}_{\hat{\beta}}(\theta)}[\ell(z; \theta)]$ obtained by drawing many i.i.d. samples from $\mathcal{D}_{\hat{\beta}}(\theta)$ (the right algorithm depends on what we know about the problem; generically we can always use DFO (Flaxman et al., 2004)). Since these samples are all synthetic and *do not* count toward the sample complexity—i.e., they do not require collecting real data but only simulation—we can draw arbitrarily many samples to achieve a small numerical error.