# Outside the Echo Chamber: Optimizing the Performative Risk

**John Miller** [*1] **Juan C. Perdomo** [*1] **Tijana Zrnic** [*1]

## Abstract

In performative prediction, predictions guide decision-making and hence can influence the distribution of future data. To date, work on performative prediction has focused on finding *performatively stable* models, which are the fixed points of repeated retraining. However, stable solutions can be far from optimal when evaluated in terms of the *performative risk*, the loss experienced by the decision maker when deploying a model. In this paper, we shift attention beyond performative stability and focus on optimizing the performative risk directly. We identify a natural set of properties of the loss function and model-induced distribution shift under which the performative risk is convex, a property which does not follow from convexity of the loss alone. Furthermore, we develop algorithms that leverage our structural assumptions to optimize the performative risk with better sample efficiency than generic methods for derivative-free convex optimization.

## 1. Introduction

Predictions in social settings are rarely made in isolation, but rather to inform decision-making. This link between predictions and decisions causes predictive models to often be *performative*, meaning they can alter their environment once deployed. For example, election forecasts impact campaign spending and affect voter turnout, hence influencing the final election outcome (Westwood et al., 2020). Similarly, long-term climate forecasts shape policy decisions which can then affect future weather patterns.

Performative prediction is a recent framework introduced by Perdomo et al. (2020) which formalizes the idea that predictive models can impact the data-generating process. So far, work in this area has focused on a particular equilibrium notion known as *performative stability* (Drusvyatskiy & Xiao, 2020; Mendler-Dünner et al., 2020; Brown et al., 2020). Stability 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 performance in the performative prediction 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 has so far remained elusive due to the complexities of model-induced distribution shift, i.e. performative effects. In particular, even in simple settings with convex losses, these distribution shifts can make the performative risk non-convex Perdomo et al. (2020). Furthermore, optimizing the performative risk requires a different algorithmic approach than what was previously studied in performative prediction. For instance, the learner needs to actively *anticipate* performative effects rather than myopically retrain until convergence, as the latter would only lead to stability.

### 1.1. Our Contributions

In this paper, we provide the first set of results describing when and how the performative risk may be optimized efficiently. We identify natural assumptions under which the performative risk is convex, even in settings where performative effects can be arbitrarily strong. Furthermore, we study optimization algorithms which explicitly model distribution shift and provably minimize the performative risk in an efficient manner.

To give an overview of our main results, we recall the relevant concepts from the performative prediction framework. Relative to supervised learning, where the learner observes

[*]Equal contribution [1]University of California, Berkeley. Correspondence to: John Miller <miller_john@berkeley.edu>, Juan C. Perdomo <jcperdomo@berkeley.edu>, Tijana Zrnic <tijana.zrnic@berkeley.edu>.

*Proceedings of the $38^{th}$ International Conference on Machine Learning*, PMLR 139, 2021. 

data from a single *static* distribution, the key conceptual innovation in the performative prediction framework is the notion of a *distribution map* $\mathcal{D}(\cdot)$, which maps model parameters $\theta \in \mathbb{R}^d$ to a distribution $\mathcal{D}(\theta)$ over instances $z$. Given a loss $\ell$, the quality of a predictive model parameterized by $\theta$ is measured according to its *performative risk*,

$$\mathrm{PR}(\theta) \stackrel{\mathrm{def}}{=} \mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta)} \ell(z;\theta).$$

A classifier $\theta_{\mathrm{PO}}$ is *performatively optimal* if it minimizes the performative risk, i.e $\theta_{\mathrm{PO}} \in \arg\min_\theta \mathrm{PR}(\theta)$. On the other hand, a classifier $\theta_{\mathrm{PS}}$ is *performatively stable* if it satisfies the fixed-point condition,

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

In other words, stable classifiers are those which are optimal for the particular distribution they induce. However, stability has little bearing on whether a classifier has low performative risk. More specifically, the following observation motivates a large part of our later analysis:

> *Stable classifiers can maximize the performative risk even when the loss is well-behaved and performative effects are small.*

Not only can stable points maximize the performative risk, but they can also have an arbitrarily large suboptimality gap, $\mathrm{PR}(\theta_{\mathrm{PS}}) - \mathrm{PR}(\theta_{\mathrm{PO}})$. Consequently, we take an alternative approach and focus on directly optimizing the performative risk.

The most natural first step towards optimizing the performative risk is to ensure that it is *convex*. Our first main result states that under an appropriate stochastic dominance condition to ensure the distribution map is well-behaved, there exists a critical threshold on the strength of performative effects which guarantees convexity:

**Theorem 1.1** (Informal)**.** *Assume that the loss is $\beta$-smooth in $z$ and $\gamma$-strongly convex in $\theta$. If the map $\mathcal{D}(\cdot)$ is $\varepsilon$-Lipschitz and satisfies a certain stochastic dominance condition, then the performative risk is guaranteed to be convex if and only if $\varepsilon \leqslant \frac{\gamma}{2\beta}$.*

Interestingly, previous work has established that $\varepsilon < \gamma/\beta$ is a threshold for repeated retraining to provably converge to a performatively stable point. Our work proves that if we halve this quantity, we get another threshold which determines whether the performative risk is convex.

While Theorem 1.1 suggests that performative effects need to be small in order to guarantee convexity, we prove that this need not be the case for the setting of *location-scale* families. These are natural classes of distribution maps in which performative effects enter through an additive or multiplicative factor that is linear in $\theta$. Many examples of distribution maps that have appeared in prior work are in fact location-scale families. For this setting, we generalize Theorem 1.1 to prove the following structural result.

**Theorem 1.2** (Informal)**.** *If the loss is smooth, strongly convex and the map $\mathcal{D}(\cdot)$ is a location-scale family, then the performative risk can be convex irrespective of the Lipschitz constant of $\mathcal{D}(\cdot)$.*

Finally, having established these structural properties, we turn to algorithms for finding performative optima. Under weak regularity assumptions, convexity alone is sufficient to apply classical zeroth-order algorithms in order to find optima in polynomial time. That said, the convergence rate of these algorithms is typically quite slow.

To address this problem, we propose a *two-stage* approach, by which the learner first creates an explicit model of the distribution map $\widehat{\mathcal{D}}$, and then optimizes a proxy objective for the performative risk obtained by "plugging in" $\widehat{\mathcal{D}}$ as if it were really the true distribution map. We instantiate this two-stage procedure in the context of location families, and prove that it optimizes the performative risk with significantly better sample efficiency then generic zeroth-order algorithms.

### 1.2. Related Work

We build on the recent line of work on performative prediction started by Perdomo et al. (2020). While previous papers in this area have focused on performative stability (Mendler-Dünner et al., 2020; Drusvyatskiy & Xiao, 2020; Brown et al., 2020), we move past this solution concept and instead analyze conditions under which one can compute performatively optimal classifiers.

Given that strategic classification is formally a special case of performative prediction (see Section 5 or discussion in Perdomo et al. (2020) for further details), the study of performative optimality has been implicitly considered in the growing body of work on strategic classification (Hardt et al., 2016; Milli et al., 2019; Hu et al., 2019; Shavit et al., 2020; Bechavod et al., 2021; Miller et al., 2020; Chen et al., 2020; Tsirtsis & Gomez Rodriguez, 2020; Haghtalab et al., 2020). 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. (2018) 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.

Similarly, several papers in strategic classification (Dong et al., 2018; Munro, 2020) and policy design (Wager & Xu, 2021) have recognized that one can apply zeroth-order algorithms (Flaxman et al., 2005; Agarwal & Dekel, 2010; Shamir, 2013) to find optima of the institution's risk. The main challenge in applying zeroth-order optimization is the fact that, in general, the performative risk might not satisfy any structural properties which would imply that its stationary points have low risk. One of the main contributions of this paper is precisely to identify under what conditions we can expect this behavior to hold.

Several works within the economics literature (Frankel & Kartik, 2021; Munro, 2020) have also contrasted fixed points of retraining and institutional optima; these analyses resemble our comparisons of stability and optimality, albeit in a more specific setting. Furthermore, there are other settings beyond strategic classification that have similarly studied optimality in the face of performative effects, such as in the context of rankings or selection bias (Rosenfeld et al., 2020; Kilbertus et al., 2020; Tabibian et al., 2020).

Lastly, our two-stage approach to minimizing the performative risk, whereby we first estimate a model of the distribution map and then optimize a proxy objective, is closely related to ideas in neighboring fields. At a high level, this general principle has appeared in semiparametric statistics (Levit, 1976; Ibragimov & Has' Minskii, 2013; Bickel, 1982; Robinson, 1988; Newey, 1990) and more recently in double machine learning (Chernozhukov et al., 2018; 2017; Mackey et al., 2018). Furthermore, this idea has been extensively studied in the controls literature where it is referred to as certainty equivalence (Theil, 1957; Simon, 1956; Mania et al., 2019; Simchowitz & Foster, 2020), or as model-based planning in reinforcement learning (Agarwal et al., 2020).

### 1.3. Additional Preliminaries

As done by previous works in this area, we limit ourselves to considering predictive models parameterized by a finite-dimensional vector $\theta \in \Theta \subseteq \mathbb{R}^d$, where $\Theta$ is a closed, convex set. The distribution map $\mathcal{D}(\cdot)$ maps parameter vectors to data distributions over real-valued instances $z \in \mathbb{R}^m$. While each model $\theta$ can induce a potentially distinct distribution $\mathcal{D}(\theta)$, we expect similar classifiers to induce similar distributions. This intuition is captured by the notion of $\varepsilon$-*sensitivity*, which is essentially a Lipschitz condition on the distribution map $\mathcal{D}(\cdot)$. We state that $\mathcal{D}(\cdot)$ is $\varepsilon$-sensitive for some $\varepsilon \geqslant 0$ if for all $\theta, \theta' \in \Theta$,

$$W_1(\mathcal{D}(\theta), \mathcal{D}(\theta')) \leqslant \varepsilon \left\|\theta - \theta'\right\|_2 . \tag{A1}$$

Here, $W_1$ denotes the Wasserstein-1 or earth mover's distance between two distributions.

## 2. Contrasting Optimality and Stability

Up until now, all works within the performative prediction literature have focused on analyzing when different algorithms converge to stable points. While the primary motivation for stability was eliminating the need for retraining, it was observed as a useful byproduct that stable points can approximately minimize the performative risk.

More specifically, Perdomo et al. (2020) prove that all stable points and performative optima lie within $\ell_2$-distance at most $2L_z\varepsilon/\gamma$ of each other, where $\varepsilon$ is the sensitivity of the distribution map, $\gamma$ denotes the strong convexity parameter of the loss, and $L_z$ denotes the Lipschitz constant of the loss *in* $z$. At first glance, this result implicitly suggests that stable points also have good predictive performance. While this is sometimes the case, in many settings $L_z$ is large enough to make the bound vacuous. For example, there exist cases where the loss function is strongly convex, but stable points actually *maximize* the performative risk.

**Proposition 2.1.** *For any $\gamma, \Delta > 0$, there exists a performative prediction problem where the loss is $\gamma$-strongly convex in $\theta$, yet the unique stable point $\theta_{\mathrm{PS}}$ 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 $\gamma$-strongly convex and the distribution map is $\varepsilon$-sensitive. A short calculation shows that the performative risk simplifies to

$$\mathrm{PR}(\theta) = \left(\frac{\gamma}{2} - \varepsilon\beta\right) \cdot \|\theta\|_2^2. \tag{1}$$

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 $\Theta$, the suboptimality gap of $\theta_{\mathrm{PS}}$ can be arbitrarily large. ■

In the above example, $\nabla_\theta \ell(z; \theta)$ is $\beta$-Lipschitz in $z$, a condition commonly referred to as *smoothness* in prior work on performativity. 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 previously studied algorithms for optimizing under performativity—such as repeated risk minimization or different variants of gradient descent (Perdomo et al., 2020; Mendler-Dünner et al., 2020)—converge to stability. Applying these methods when $\varepsilon \in (\gamma/(2\beta), \gamma/\beta)$ would hence maximize the performative risk on this problem.

Moreover, we remark that the Lipschitz constant $L_z$ is equal to $\beta \cdot \max_{\theta\in\Theta} \|\theta\|_2$. Therefore, the results of Perdomo et al. (2020) imply that stable points and optima are at distance at most $\frac{2L_z\varepsilon}{\gamma} = \frac{2\beta\varepsilon}{\gamma} \max_{\theta\in\Theta} \|\theta\|_2$. When $\varepsilon > \frac{\gamma}{2\beta}$, as assumed in the proof of Proposition 2.1, this bound on the distance becomes vacuous: $\|\theta_{\mathrm{PS}} - \theta_{\mathrm{PO}}\|_2 \leqslant \max_{\theta\in\Theta} \|\theta\|_2$.

Lastly, we 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 (1). 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.

## 3. Convexity of the Performative Risk

We now introduce our main 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 $\lambda$-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 $\lambda$ is positive, then $\mathrm{PR}(\theta)$ is $\lambda$-strongly convex. If $\lambda$ is negative, then adding the analogous regularizer $\frac{\lambda}{2}\|\theta\|^2$ ensures $\mathrm{PR}(\theta)$ is convex. Furthermore, in addition to $\varepsilon$-sensitivity, we will make repeated use of the following assumptions throughout the remainder of the paper. To facilitate readability, we let $\mathcal{Z} \stackrel{\mathrm{def}}{=} \cup_{\theta\in\Theta}\mathrm{supp}(\mathcal{D}(\theta))$.

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

$$\|\nabla_\theta \ell(z;\theta) - \nabla_\theta \ell(z';\theta)\|_2 \leqslant \beta \|z - z'\|_2 . \tag{A2}$$

Furthermore, a loss function $\ell(z;\theta)$ is $\gamma$-*strongly convex* in $\theta$ if for all $\theta, \theta', \theta'' \in \Theta$,

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta'')} \ell(z;\theta) \geqslant \mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta'')} \ell(z;\theta') + \mathop{\mathbb{E}}_{z\sim\mathcal{D}(\theta'')} \nabla_\theta \ell(z;\theta')^\top(\theta - \theta') + \frac{\gamma}{2}\|\theta - \theta'\|_2^2 . \tag{A3a}$$

If $\gamma = 0$, this assumption is equivalent to convexity. Similarly, we say that the loss is $\gamma_z$-strongly convex in $z$ if for all $\theta \in \Theta$ and $z, z' \in \mathcal{Z}$,

$$\ell(z;\theta) \geqslant \ell(z';\theta) + \nabla_z \ell(z';\theta)^\top(z' - z) + \frac{\gamma_z}{2}\|z - z'\|_2^2 . \tag{A3b}$$

Lastly, we state that a distribution map, loss pair $(\mathcal{D}(\cdot), \ell)$ satisfies *mixture dominance* if the following condition holds for all $\theta, \theta', \theta'' \in \Theta$ and $\alpha \in (0, 1)$:

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}(\alpha\theta+(1-\alpha)\theta')} \ell(z;\theta'') \leqslant \mathop{\mathbb{E}}_{z\sim\alpha\mathcal{D}(\theta)+(1-\alpha)\mathcal{D}(\theta')} \ell(z;\theta'') \tag{A4}$$

Smoothness and strong convexity are standard and have appeared previously in the context of performative prediction. The mixture dominance condition is novel and plays a central role in our analysis of when the performative risk is convex. To provide some intuition for this condition, we recall the definition of the *decoupled performative risk*:

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

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 (A3a) guarantees that DPR is convex in the second argument, mixture dominance (A4) 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 (A4) 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 (Shaked & Shanthikumar, 2007), which we further discuss in Appendix A. 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 (A4) holds when the loss is convex in $z$ and the distribution map $\mathcal{D}(\cdot)$ forms a *location-scale family* of the form:

$$z_\theta \sim \mathcal{D}(\theta) \;\Leftrightarrow\; z_\theta \stackrel{d}{=} (\Sigma_0 + \Sigma(\theta))z_0 + \mu_0 + \mu\theta, \tag{2}$$

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 (see Proposition A.4 for a formal proof). Distribution maps of this sort are ubiquitous throughout the performative prediction literature and hence satisfy mixture dominance if the loss $\ell$ is convex. For instance, the distribution map for the strategic classification simulator in Perdomo et al. (2020) is a location family. Other examples of location families can be found in previous work on strategic classification (Frankel & Kartik, 2021; Haghtalab et al., 2020). Mixture dominance can also hold in discrete settings, e.g. $\mathcal{D}(\theta) = \mathrm{Bernoulli}(a^\top\theta + b)$ 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 3.1.** *Suppose that the loss function $\ell(z;\theta)$ is $\gamma$-strongly convex in $\theta$ (A3a), $\beta$-smooth in $z$ (A2), and that $\mathcal{D}(\cdot)$ is $\varepsilon$-sensitive (A1). If mixture dominance (A4) holds, then the performative risk is $\lambda$-convex for $\lambda = \gamma - 2\varepsilon\beta$.*

Together with the example from the proof of Proposition 2.1, this theorem shows that $\frac{\gamma}{2\beta}$ is a sharp threshold for convexity of the performative risk. If $\varepsilon$ is strictly less than this

threshold, then under mixture dominance and appropriate conditions on the loss, the performative risk is strongly convex by Theorem 3.1. On the other hand, if $\varepsilon$ is above this threshold, the example from Proposition 2.1 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. This threshold was also implicitly observed by Perdomo et al. (2020) in the proof of Proposition 4.2 as byproduct of showing that the performative risk can be non-convex for $\varepsilon \leqslant \frac{\gamma}{\beta}$. However, they provide no general analysis of when the performative risk is convex. Note that all of the above examples satisfy mixture dominance.

While the threshold $\varepsilon = \gamma/(2\beta)$ is in general tight as argued above, for certain families of distribution maps the conclusion of Theorem 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 3.2.** 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 $\xi$ is a zero-mean noise variable. If $\ell$ 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 $\mu$.

The proof follows by simply observing that in both cases, the performative risk can be written as a linear function in $\theta$ composed with a convex function. Another interesting property of this example is that the distribution map is $\varepsilon$-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 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 3.3.** *Suppose that $\ell(z; \theta)$ is $\gamma$-strongly convex in $\theta$ (A3a), $\beta$-smooth (A2), and $\gamma_z$-strongly convex in $z$ (A3b). Furthermore, suppose that $\mathcal{D}(\theta)$ forms a location-scale family (2) with $\varepsilon$ as its sensitivity parameter*[1]*. Define $\Sigma_{z_0}$ 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 $\lambda$-convex for $\lambda$ equal to:*

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

[1]The sensitivity parameter $\varepsilon$ for location-scale families can be explicitly bounded in terms of the parameters $\mu$ and $\Sigma(\theta)$; see Remark C.3 in the Appendix.

**Remark 3.4.** *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 performativity. 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 $\theta$, and performative variables which are influenced by the deployed classifier. For example, in strategic classification, the performative effects are often only present in the strategically manipulated features, and not in the label. In Example 3.2, on the other hand, the effects are only present in the label. For simplicity of exposition, we suppress this distinction between* performative *and* static *variables, that is, those whose distribution does not change for different $\mathcal{D}(\theta)$. However, the reader should think of all assumptions on $z$, such as strong convexity or various Lipschitz assumptions, as only having to apply to the performative variables, while the static ones can be averaged out. We elaborate on how the analysis can be strengthened in Appendix B.*

We now illustrate an application of Theorem 3.3 on a scale family example.

**Example 3.5.** Suppose that $x > 0$ is a one-dimensional feature drawn from a fixed distribution $\mathcal{D}_x$ with finite second moment, and let $y|x \sim \theta x \cdot \mathrm{Exp}(1)$ be distributed as an exponential random variable with mean $\theta x$. 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 $\theta$ and $\mathbb{E}\, x$-smooth in $y$. Therefore, according to Theorem 3.3, the performative risk is convex and hence tractable to optimize, since $\gamma - \beta^2/\gamma_z = \mathbb{E}\, x^2 - (\mathbb{E}\, x)^2 \geqslant 0$ 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 2.1 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 3.3, we provide another scale family example where the performative risk is convex regardless of $\varepsilon$, yet stable points can be arbitrarily suboptimal.

**Example 3.6.** 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 $\varepsilon$-sensitive. If $\ell$ is the squared loss, $\ell(z;\theta) = \frac{1}{2}(z-\theta)^2$, then there is a unique stable point $\theta_{\mathrm{PS}} = \mu$. We remark how stability is completely oblivious to the performative effects, as $\arg\min_{\theta'} \mathbb{E}_{z\sim\mathcal{D}(\theta)} \ell(z;\theta') = \mu$ regardless of $\theta$. The performative optimum is $\theta_{\mathrm{PO}} = \mu/(1+\varepsilon^2)$. Depending on $\mu$, 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 3.3, 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.

## 4. Optimization Algorithms

Having identified conditions under which the performative risk is convex, we now consider methods for efficiently optimizing it. One of the main challenges of carrying out this task is that, even in convex settings, the learner can only access the objective via noisy function evaluations corresponding to 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 $\varepsilon$ (as indeed confirmed by Theorem 3.3, 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 $p_\theta$, 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 in performative prediction, such as repeated retraining (Perdomo et al., 2020), fail on this problem. The reason is that they essentially ignore the second term in the gradient computation which requires explicitly anticipating performative effects. For example, retraining computes the sequence of updates $\theta_{t+1} = \arg\min_\theta \mathbb{E}_{z\sim\mathcal{D}(\theta_t)} \frac{1}{2}(z-\theta)^2 = \varepsilon\theta_t$, which diverges for $|\varepsilon| > 1$.

### 4.1. 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 (Flaxman et al., 2005; Agarwal & Dekel, 2010; Shamir, 2013). 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. (2005) 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 $\Delta$-suboptimal point (Agarwal & Dekel, 2010; Shamir, 2013), which can be prohibitively expensive if samples are hard to come by.

### 4.2. 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

**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 OLS, $\widehat{\mu} \in \arg\min_\mu \sum_{i=1}^n \|z_i - \mu\theta_i\|_2^2$.
// 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, $\arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{j=n+1}^{2n} \ell(z_j + \widehat{\mu}\theta; \theta)$.

the estimated $\widehat{\mathcal{D}}$ as if it were the true distribution map:

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

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 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.[2]

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 $\mu$ which captures the performative effects. In particular, we sample and deploy $n$ classifiers $\theta_i$, $i \in [n]$, observe data $z_i \sim \mathcal{D}(\theta_i)$, and then construct an estimate $\widehat{\mu}$ of the location map $\mu$ 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 $\theta'$, we can simulate $z \sim \widehat{\mathcal{D}}(\theta')$ by taking $z = z_0 + \widehat{\mu}\theta'$.

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

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

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

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

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 $\Delta$-suboptimal classifier requires $O(d/\Delta)$ samples. In contrast, the derivative-free methods considered previously require $\widetilde{O}(d^2/\Delta^2)$ samples to compute a classifier of similar quality. The formal statement and proof of this result is deferred to Appendix C.2.

**Theorem 4.1** (Informal)**.** *Under appropriate smoothness and strong convexity assumptions on the loss $\ell$, if the distribution of $z_0$ is subgaussian, and if the number of samples $n \geqslant \Omega(d + m + \log(1/\delta))$, then, with probability $1-\delta$, Algorithm 1 returns a point $\widehat{\theta}_n$ such that*

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

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.

## 5. Experiments

We complement our theoretical findings with an empirical evaluation of different methods on two tasks: the strategic classification simulator from Perdomo et al. (2020), 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 4.1, versus

[2] 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.

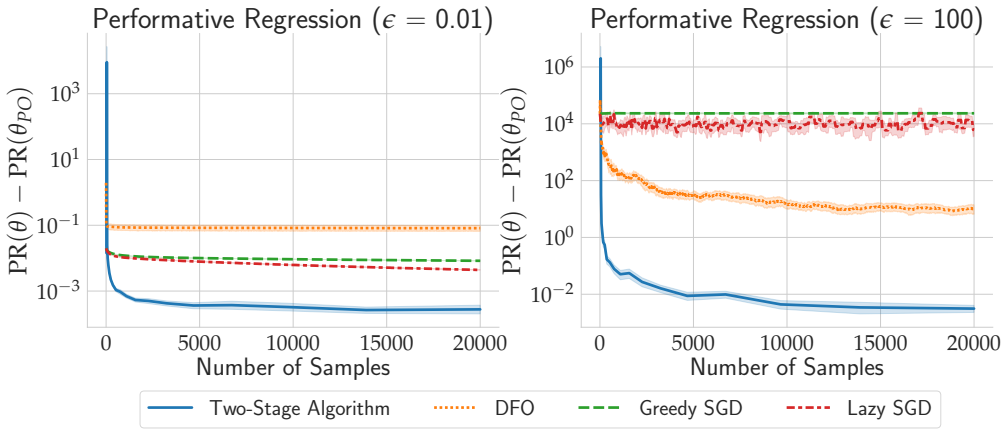


*Figure 1.* Suboptimality gap versus number of samples collected for the two-stage algorithm, DFO algorithm, greedy SGD, and lazy SGD, for $\varepsilon = 0.01$ (left) and $\varepsilon = 100$ (right).

existing optimization algorithms for finding stable points, in particular greedy and lazy SGD due to Mendler-Dünner et al. (2020). 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 3. To evaluate derivative-free methods, we implement the "gradient descent without a gradient" algorithm from Flaxman et al. (2005), which we refer to from here on out as the "DFO algorithm." For each of the following experiments, we run each algorithm 50 times and display 95% bootstrap confidence intervals. We provide a formal description of all the procedures, as well as a detailed description of the experimental setup in Appendix D.

### 5.1. Linear Regression

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 3.2, where the learner attempts to solve a linear regression with performative labels. Given a parameter $\theta$, 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 $\varepsilon$-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 $\mu$.

For small $\varepsilon$, greedy and lazy SGD converge to a stable point that approximately minimizes the performative risk (see left panel in Figure 1). However, as the strength of performative effects increases, these methods fail to make progress and are outperformed by the DFO algorithm and the two-stage approach by a considerable margin (see right panel in Figure 1). The two-stage procedure efficiently converges after a small number of samples and its behavior is largely unaffected as we increase the value of $\varepsilon$, while the DFO algorithm becomes considerably slower when $\varepsilon$ is large.

### 5.2. Strategic Classification Simulator

We next consider experiments on the credit scoring simulator from Perdomo et al. (2020), which has been employed as an empirical benchmark for performative prediction in several works (Mendler-Dünner et al., 2020; Drusvyatskiy & Xiao, 2020; Brown et al., 2020). The simulator models a strategic classification problem between a bank and individual agents seeking a loan. The bank deploys a logistic regression classifier $f_\theta$ to determine the individuals' default probabilities, while agents strategically manipulate their features to achieve a more favorable classification.

More specifically, individuals correspond to feature, label pairs $(x, y)$ drawn i.i.d. from a base distribution $\mathcal{D}_0$. Given a classifier $f_\theta$, agents compute a best-response set of features $x_{\mathrm{BR}}$ by solving an optimization problem. The bank then observes the manipulated data points $(x_{\mathrm{BR}}, y) \sim \mathcal{D}(\theta)$. For an appropriate choice of the agents' objective function, the distribution map forms a location family, $x_{\mathrm{BR}} = x + \varepsilon\theta$, where $\varepsilon$ is a parameter of the agents' objective. It also serves as a measure of performativity, since this distribution map is $\varepsilon$-sensitive. As a final remark, we add $\ell_2$-regularization to the logistic loss to ensure strong convexity. See Perdomo et al. (2020) and Appendix D for full details.

Since the logistic loss is not strongly convex in the features, we only have a certificate of convexity when $\varepsilon$ is small enough (namely, $\varepsilon \leqslant \frac{\gamma}{2\beta}$). We consider two values of $\varepsilon$: one which is below this critical threshold, and one large value for which we do not have theoretical guarantees.

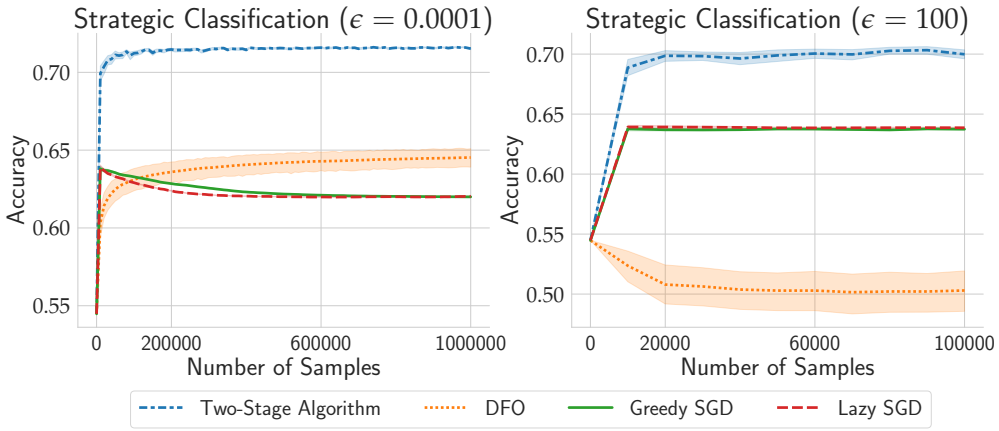


*Figure 2.* Classification accuracy versus number of samples collected for the two-stage algorithm, DFO algorithm, greedy SGD, and lazy SGD, for $\varepsilon = 0.0001 \leqslant \frac{\gamma}{2\beta}$ (left) and $\varepsilon = 100 \gg \frac{\gamma}{2\beta}$ (right).

When $\varepsilon$ 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). 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.

## 6. Discussion and Future Work

Given the stark difference between performative stability and optimality, the goal of our work was to identify the first set of conditions and algorithmic procedures by which one might be able to provably optimize the performative risk. To this end, we focused on analyzing the problem at a broad level of generality, identifying simple, structural conditions under which the optimization problem becomes tractable.

However, when applying these ideas in practice, there are a number of important considerations determined by the relevant social context that are not explicitly addressed by our theoretical analysis and which we believe are an important direction for future work. For example, in social domains such as credit scoring or election forecasts, the choice of loss function must balance predictive accuracy with any possible externalities of the impact of classifier deployment (i.e. performative effects) on the observed distribution. In a similar vein, exploration strategies must be weighed against the relevant costs of deploying a particular model. We believe that elucidating these tradeoffs and design choices in the context of specific applications and case studies on performative prediction would be of high value to the community.

## Acknowledgements

We thank Moritz Hardt and Celestine Mendler-Dünner for many helpful conversations during the course of this project as well as for providing detailed feedback on a draft of this manuscript.

This research was generously supported in part by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 1752814.

## References

Agarwal, A. and Dekel, O. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Conference on Learning Theory*, pp. 28–40, 2010.

Agarwal, A., Kakade, S., and Yang, L. F. Model-based rein-

forcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pp. 67–83, 2020.

Bechavod, Y., Ligett, K., Wu, S., and Ziani, J. Gaming helps! learning from strategic interactions in natural dynamics. In *International Conference on Artificial Intelligence and Statistics*, pp. 1234–1242, 2021.

Bickel, P. J. On adaptive estimation. *The Annals of Statistics*, pp. 647–671, 1982.

Brown, G., Hod, S., and Kalemaj, I. Performative prediction in a stateful world. *arXiv preprint arXiv:2011.03885*, 2020.

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

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. Double/debiased/Neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, 2017.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters, 2018.

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. ACM, 2018.

Drusvyatskiy, D. and Xiao, L. Stochastic optimization with decision-dependent distributions. *arXiv preprint arXiv:2011.11173*, 2020.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Symposium on Discrete Algorithms*, pp. 385–394, 2005.

Frankel, A. and Kartik, N. Improving information from manipulable data. *Journal of the European Economic Association*, 2021.

Haghtalab, N., Immorlica, N., Lucier, B., and Wang, J. Z. Maximizing welfare with incentive-aware evaluation mechanisms. In *International Joint Conference on Artificial Intelligence*, 2020.

Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. Strategic classification. In *Proceedings of the ACM Conference on Innovations in Theoretical Computer Science*, pp. 111–122, 2016.

Hu, L., Immorlica, N., and Vaughan, J. W. The disparate effects of strategic manipulation. In *Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency*, pp. 259–268, 2019.

Ibragimov, I. A. and Has' Minskii, R. Z. *Statistical estimation: asymptotic theory*, volume 16. Springer Science & Business Media, 2013.

Kilbertus, N., Rodriguez, M. G., Schölkopf, B., Muandet, K., and Valera, I. Fair decisions despite imperfect predictions. In *International Conference on Artificial Intelligence and Statistics*, pp. 277–287, 2020.

Levit, B. Y. On the efficiency of a class of non-parametric estimates. *Theory of Probability & Its Applications*, 20 (4):723–740, 1976.

Mackey, L., Syrgkanis, V., and Zadik, I. Orthogonal machine learning: Power and limitations. In *International Conference on Machine Learning*, pp. 3375–3383, 2018.

Mania, H., Tu, S., and Recht, B. Certainty equivalence is efficient for linear quadratic control. In *Advances in Neural Information Processing Systems*, pp. 10154–10164, 2019.

Matni, N. and Tu, S. A tutorial on concentration bounds for system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 3741–3749. IEEE, 2019.

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

Miller, J., Milli, S., and Hardt, M. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pp. 6917–6926. PMLR, 2020.

Milli, S., Miller, J., Dragan, A. D., and Hardt, M. The social cost of strategic classification. In *Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency*, pp. 230–239, 2019.

Müller, A. and Rüschendorf, L. On the optimal stopping values induced by general dependence structures. *Journal of applied probability*, pp. 672–684, 2001.

Müller, A. and Stoyan, D. *Comparison methods for stochastic models and risks*, volume 389. Wiley, 2002.

Munro, E. Learning to personalize treatments when agents are strategic. *arXiv preprint arXiv:2011.06528*, 2020.

Newey, W. K. Semiparametric efficiency bounds. *Journal of applied econometrics*, 5(2):99–135, 1990.

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

Robinson, P. M. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pp. 931–954, 1988.

Roger, H. and Charles, R. J. Topics in matrix analysis, 1994.

Rosenfeld, N., Hilgard, A., Ravindranath, S. S., and Parkes, D. C. From predictions to decisions: Using lookahead regularization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4115–4126, 2020.

Ross, S. M., Kelly, J. J., Sullivan, R. J., Perry, W. J., Mercer, D., Davis, R. M., Washburn, T. D., Sager, E. V., Boyce, J. B., and Bristow, V. L. *Stochastic processes*, volume 2. Wiley New York, 1996.

Shaked, M. and Shanthikumar, J. G. *Stochastic orders*. Springer Science & Business Media, 2007.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

Shamir, O. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory*, pp. 3–24, 2013.

Shavit, Y., Edelman, B., and Axelrod, B. Causal strategic linear regression. In *International Conference on Machine Learning*, pp. 8676–8686, 2020.

Simchowitz, M. and Foster, D. Naive exploration is optimal for online LQR. In *International Conference on Machine Learning*, pp. 8937–8948, 2020.

Simon, H. A. Dynamic programming under uncertainty with a quadratic criterion function. *Econometrica, Journal of the Econometric Society*, pp. 74–81, 1956.

Tabibian, B., Gomez, V., De, A., Schölkopf, B., and Rodriguez, M. G. On the design of consequential ranking algorithms. In *Conference on Uncertainty in Artificial Intelligence*, pp. 171–180, 2020.

Theil, H. A note on certainty equivalence in dynamic planning. *Econometrica: Journal of the Econometric Society*, pp. 346–349, 1957.

Tsirtsis, S. and Gomez Rodriguez, M. Decisions, counterfactual explanations and strategic behavior. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16749–16760, 2020.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Wager, S. and Xu, K. Experimenting in equilibrium. *Management Science*, 2021.

Westwood, S. J., Messing, S., and Lelkes, Y. Projecting confidence: How the probabilistic horse race confuses and demobilizes the public. *The Journal of Politics*, 82 (4):1530–1544, 2020.