Performative Prediction

Juan C. Perdomo [*1] **Tijana Zrnic** [*1] **Celestine Mendler-Dünner** [1] **Moritz Hardt** [1 2]

Abstract

When predictions support decisions they may influence the outcome they aim to predict. We call such predictions *performative*; the prediction influences the target. Performativity is a well-studied phenomenon in policy-making that has so far been neglected in supervised learning. When ignored, performativity surfaces as undesirable distribution shift, routinely addressed with retraining. We develop a risk minimization framework for performative prediction bringing together concepts from statistics, game theory, and causality. A conceptual novelty is an equilibrium notion we call performative stability. Performative stability implies that the predictions are calibrated not against past outcomes, but against the future outcomes that manifest from acting on the prediction. Our main results are necessary and sufficient conditions for the convergence of retraining to a performatively stable point of nearly minimal loss. In full generality, performative prediction strictly subsumes the setting known as *strategic classification*. We thus also give the first sufficient conditions for retraining to overcome strategic feedback effects.

1. Introduction

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

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

Once recognized, performativity turns out to be ubiquitous. Traffic predictions influence traffic patterns, crime location prediction influences police allocations that may deter crime, recommendations shape preferences and thus consumption, stock price prediction determines trading activity and hence prices. When ignored, performativity can surface as a form of *distribution shift*. As the decision-maker acts according to a predictive model, the distribution over data points appears to change over time. In practice, the response to such distribution shifts is to frequently *retrain* the predictive model as more data becomes available. Retraining is often considered an undesired—yet necessary—cat and mouse game of chasing a moving target.

What would be desirable from the perspective of the decision maker is a certain equilibrium where the model is optimal for the distribution it induces. Such equilibria coincide with the stable points of retraining, that is, models invariant under retraining. Performativity therefore suggests a different perspective on retraining, exposing it as a natural equilibrating dynamic rather than a nuisance.

This raises fundamental questions. When do such stable points exist? How can we efficiently find them? Under what conditions does retraining converge? When do stable points also have good predictive performance? In this work, we formalize performative prediction, tying together conceptual elements from statistical decision theory, causal reasoning, and game theory. We then resolve some of the fundamental questions that performativity raises.

1.1. Our Contributions

We put performativity at the center of a decision-theoretic framework that extends the classical statistical theory underlying risk minimization. The goal of risk minimization is to find a decision rule, specified by model parameters θ, that performs well on a fixed joint distribution $\mathcal{D}$ over covariates X and an outcome variable Y. Whenever predictions are performative, the choice of predictive model affects the observed distribution over instances $Z = (X, Y)$. We formalize this intuitive notion by introducing a map $\mathcal{D}(\cdot)$ from

[*]Equal contribution [1]University of California, Berkeley [2]MH is a paid consultant for Twitter. Correspondence to: Juan C. Perdomo <jcperdomo@berkeley.edu>, Tijana Zrnic <tijana.zrnic@berkeley.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

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

A natural objective in performative prediction is to evaluate model parameters θ on the resulting distribution $\mathcal{D}(\theta)$ as measured via a loss function ℓ. This results in the notion we call *performative risk*, defined as

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

The difficulty in minimizing $\mathrm{PR}(\theta)$ is that the distribution itself depends on the argument θ, a dependence that defeats traditional theory for risk minimization. Moreover, we generally envision that the map $\mathcal{D}(\cdot)$ is unknown to the decision maker.

Perhaps the most natural algorithmic heuristic in this situation is a kind of fixed point iteration: repeatedly find a model that minimizes risk on the distribution resulting from the previous model, corresponding to the update rule

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

We call this procedure *repeated risk minimization*. We also analyze its empirical counterpart, where we work with finite samples. These procedures exemplify a family of *retraining* heuristics that are ubiquitous in practice for dealing with all kinds of distributions shifts irrespective of cause.

When repeated risk minimization converges in objective value, the model has minimal loss on the distribution it entails:

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

We refer to this condition as *performative stability*, noting that it is neither implied by nor does it imply minimal performative risk. Our central result can be summarized informally as follows.

Theorem 1.1 (Informal)**.** *If the loss is smooth, strongly convex, and the map $\mathcal{D}(\cdot)$ is sufficiently Lipschitz, then repeated risk minimization converges to performative stability at a linear rate. Moreover, if any one of these assumptions does not hold, repeated risk minimization can fail to converge.*

The notion of Lipschitz continuity here refers to the Euclidean distance on model parameters and the earth mover's distance on distributions. Informally, it requires that a small change in model parameters θ does not have an outsized effect on the induced distribution $\mathcal{D}(\theta)$.

In contrast to standard supervised learning, convexity alone is *not* sufficient for convergence in objective value, even if the other assumptions hold. Performative prediction therefore gives a new and interesting perspective on the importance of strong convexity.

Strong convexity has a second benefit. Not only does it guarantee that retraining converges to a stable point at a linear rate, it also ensures that this stable point approximately minimizes the performative risk.

Theorem 1.2 (Informal)**.** *If the loss is Lipschitz and strongly convex, and the map $\mathcal{D}(\cdot)$ is Lipschitz, all stable points and performative optima lie in a small neighborhood.*

Recall that performative stability on its own does not imply minimal performative risk. What the previous theorem shows, however, is that strong convexity guarantees that we can approximately satisfy both.

We complement our main results with a case study in *strategic classification*. Strategic classification aims to anticipate a strategic response to a classifier from an individual, who can change their features prior to being classified. We observe that strategic classification is a special case of performative prediction. On the one hand, this allows us to transfer our technical results to this established setting. In particular, our results are the first to give a guarantee on repeated risk minimization in the strategic setting. On the other hand, strategic classification provides us with one concrete setting for what the mapping $\mathcal{D}(\cdot)$ can be. We use this as a basis of an empirical evaluation in a semi-synthetic setting, where the initial distribution is based on a real data set, but the distribution map is modeled.

1.2. Related Work

Performativity is a broad concept in the social sciences, philosophy, and economics (MacKenzie et al., 2007; Healy, 2015). Below we focus on the relationship of our work to the most relevant technical scholarship.

Learning on non-stationary distributions. A closely related line of work considers the problem of *concept drift*, broadly defined as the problem of learning when the target distribution over instances drifts with time (Kuh et al., 1991; Bartlett, 1992; Bartlett et al., 2000; Gama et al., 2014).

Concept drift is more general phenomenon than performativity in that it considers arbitrary sources of shift. However, studying the problem at this level of generality has led to a number of difficulties in creating a unified language and objective (Gama et al., 2014; Webb et al., 2016), an issue we circumvent by assuming that the population distribution is determined by the deployed classifier. Importantly, this line of work also discusses the importance of retraining (Žliobaitė, 2010; Gama et al., 2014). However, it stops short of discussing the need for stability or analyzing the long-term behavior of retraining.

Strategic classification. Strategic classification recognizes that individuals often adapt to the specifics of a decision rule so as to gain an advantage (Dalvi et al., 2004; Brückner et al., 2012; Hardt et al., 2016a; Khajehnejad et al., 2019). Recent work in this area considers issues of incentive design (Kleinberg & Raghavan, 2019; Miller et al., 2020), control over an algorithm (Burrell et al., 2019), and fairness (Hu et al., 2019; Milli et al., 2019). Our model of performative prediction includes all notions of strategic adaption that we are aware of as a special case. Unlike many works in this area, our results do not depend on a specific *cost function* for changing individual features. Rather, we rely on an assumption about the sensitivity of the data-generating distribution to changes in the model parameters.

Recently, there has been increased interest within the algorithmic fairness community in classification dynamics. See, for example, Liu et al. (2018), Hu & Chen (2018), and Hashimoto et al. (2018). The latter work considers repeated risk minimization, but from the perspective of what it does to a measure of disparity between groups.

Causal inference. The reader familiar with causality can think of $\mathcal{D}(\theta)$ as the interventional distribution over instances Z resulting from a do-intervention that sets the model parameters to θ in some underlying causal graph. Importantly, this mapping $\mathcal{D}(\cdot)$ remains fixed and does not change over time or by intervention: deploying the same classifier at two different points in time must induce the same distribution over observations Z. While causal inference focuses on estimating properties of interventional distributions such as treatment effects (Pearl, 2009; Imbens & Rubin, 2015), our focus is on a new stability notion and iterative retraining procedures for finding stable points.

Reinforcement learning. In general, any instance of performative prediction can be reframed as a reinforcement learning problem. Yet, by studying performative prediction problems within such a broad framework, we lose many of the intricacies of performativity which make the problem interesting and tractable to analyze. We return to discuss some of the connections between both frameworks later on.

2. Framework and Main Definitions

In this section, we formally introduce the principal solution concepts of our framework: performative optimality and performative stability.

Throughout our presentation, we focus on predictive models that are parametrized by a vector $\theta \in \Theta$, where the parameter space $\Theta \subseteq \mathbb{R}^d$ is a closed, convex set. We use capital letters to denote random variables and their lowercase counterparts to denote realizations of these variables. We consider instances $z = (x, y)$ defined as feature, outcome pairs, where $x \in \mathbb{R}^{m-1}$ and $y \in \mathbb{R}$. Whenever we define a variable $\theta^* = \arg\min_\theta g(\theta)$ as the minimizer of a function g, we resolve the issue of the minimizer not being unique by setting θ^* to an arbitrary point in the $\arg\min_\theta g(\theta)$ set.

2.1. Performative Optimality

In supervised learning, the goal is to learn a predictor f_θ which minimizes the expected loss with respect to instances drawn i.i.d. from a fixed distribution $\mathcal{D}$. The optimal classifier $f_{\theta_{\mathrm{SL}}}$ solves the following optimization problem,

$$\theta_{\mathrm{SL}} = \arg\min_{\theta \in \Theta} \mathop{\mathbb{E}}_{Z \sim \mathcal{D}} \ell(Z; \theta),$$

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

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

Definition 2.1 (performative optimality and risk)**.** *A classifier $f_{\theta_{\mathrm{PO}}}$ is* performatively optimal *if the following relationship holds:*

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

We define $\mathrm{PR}(\theta) \stackrel{\mathrm{def}}{=} \mathbb{E}_{Z \sim \mathcal{D}(\theta)} \ell(Z; \theta)$ *as the* performative risk*; then,* $\theta_{\mathrm{PO}} = \arg\min_\theta \mathrm{PR}(\theta)$.

The following example illustrates the differences between the traditional notion of optimality in supervised learning and performative optima. Appendix C contains full derivations relevant to this example.

Example 2.2 (biased coin flip)**.** *Consider the task of predicting a biased coin flip where the bias of the coin depends on a feature X and the assigned score $f_\theta(X)$.*

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

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

In the performative setting with $\varepsilon \neq 0$, the optimal prediction θ_{PO} balances between its predictive accuracy as well

as the bias induced by the prediction itself. In particular, a direct calculation demonstrates that

$$\theta_{\mathrm{PO}} = \arg\min_{\theta\in[0,1]} \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta)} \left(Y - \theta X - \frac{1}{2}\right)^2 \Leftrightarrow \theta_{\mathrm{PO}} = \frac{\mu}{1-2\varepsilon}.$$

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

2.2. Performative Stability

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

Definition 2.3 (performative stability and decoupled risk)**.** *A classifier $f_{\theta_{\mathrm{PS}}}$ is* performatively stable *if the following relationship holds:*

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

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

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

Observe that performative optimality and performative stability are in general two distinct solution concepts. Performatively optimal classifiers need not be performatively stable and performatively stable classifiers need not be performatively optimal. We illustrate this point in the context of our previous biased coin toss example.

Example 2.2 (continued). *Consider again our model of a biased coin toss. In order for a classifier f_θ to be performatively stable, it must satisfy the following relationship:*

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

Solving for θ_{PS} directly, we see that there is a unique performatively stable classifier. This example illustrates that performative stability and performative optimality need not identify. In fact, in this example they identify if and only if $\varepsilon = 0$. Note that, in general, if the map $\mathcal{D}(\theta)$ is constant across θ, performative optima must coincide with performatively stable solutions, and both coincide with "static" supervised learning solutions as well.

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

Remark 2.4. *Observe that both performative stability and optimality can be expressed via the decoupled performative risk as follows:*

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

3. When Retraining Converges to Stable Points

Having introduced our framework for performative prediction, we now address some of the basic questions that arise in this setting and examine the behavior of common machine learning practices, such as retraining, through the lens of performativity. We begin by analyzing the behavior of these procedures when they operate at a population level and then extend our analysis to finite samples.

3.1. Assumptions

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

Definition 3.1 (ε-sensitivity)**.** *We say that a distribution map $\mathcal{D}(\cdot)$ is* ε-sensitive *if for all $\theta,\theta' \in \Theta$:*

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

where W_1 denotes the earth mover's distance.

The earth mover's distance is a natural notion of distance between probability distributions that provides access to a rich technical repertoire (Villani, 2003; 2008). Furthermore, we can verify that it is satisfied in various settings.

Remark 3.2. *An example where this assumption is satisfied is a Gaussian family: given $\theta = (\mu,\sigma_1,\ldots,\sigma_p) \in \mathbb{R}^{2p}$, define $\mathcal{D}(\theta) = \mathcal{N}(\varepsilon_1\mu, \varepsilon_2^2\,\mathrm{diag}(\sigma_1^2,\ldots,\sigma_p^2))$ where $\varepsilon_1,\varepsilon_2 \in \mathbb{R}$. Then $\mathcal{D}(\cdot)$ is ε-sensitive for $\varepsilon = \max\big\{|\varepsilon_1|, |\varepsilon_2|$.*

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

(A1) (joint smoothness) A loss function $\ell(z;\theta)$ is β-jointly smooth if $\forall\theta,\theta' \in \Theta$ and $z, z' \in \mathcal{Z}$:

$$\|\nabla_\theta \ell(z;\theta) - \nabla_\theta \ell(z;\theta')\|_2 \leqslant \beta \|\theta - \theta'\|_2 ,$$
$$\|\nabla_\theta \ell(z;\theta) - \nabla_\theta \ell(z';\theta)\|_2 \leqslant \beta \|z - z'\|_2 .$$

(A2) (strong convexity) A loss function $\ell(z;\theta)$ is γ-strongly convex if $\forall\,\theta,\theta' \in \Theta$ and $z \in \mathcal{Z}$:

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

If $\gamma = 0$, this last assumption is equivalent to convexity. We will sometimes refer to $\frac{\beta}{\gamma}$, where β is defined as in (A1) and γ as in (A2), as the condition number.

3.2. Repeated Risk Minimization

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

Definition 3.3 (RRM)**.** Repeated risk minimization *(RRM) refers to the procedure where, starting from an initial model f_{θ_0} we perform the following sequence of updates for $t \geqslant 0$:*

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

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

Example 3.4. *Consider optimizing the squared loss $\ell(z;\theta) = (y-\theta)^2$, where $\theta \in [0,1]$ and the distribution of the outcome Y, according to $\mathcal{D}(\theta)$, is a point mass at 0 if $\theta \geqslant \frac{1}{2}$, and a point mass at 1 if $\theta < \frac{1}{2}$. Clearly there is no performatively stable point, and RRM will simply result in the alternating sequence $1, 0, 1, 0, \ldots$. The performative optimum in this case is $\theta_{\text{PO}} = \frac{1}{2}$.*

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

Theorem 3.5. *Suppose that the loss $\ell(z;\theta)$ is β-jointly smooth (A1) and γ-strongly convex (A2). If the distribution map $\mathcal{D}(\cdot)$ is ε-sensitive, then the following is true:*

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

(b) If $\varepsilon < \frac{\gamma}{\beta}$, the iterates θ_t of RRM converge to a unique performatively stable point θ_{PS} at a linear rate,

$$\|\theta_t - \theta_{\text{PS}}\|_2 \leqslant \delta \text{ for } t \geqslant \frac{\log(\frac{1}{\delta}\|\theta_0 - \theta_{\text{PS}}\|_2)}{(1-\varepsilon\frac{\beta}{\gamma})}.$$

The main message of this theorem is that in performative prediction, if the loss function is sufficiently "nice" and the distribution map is sufficiently (in)sensitive, then one need only retrain a classifier a small number of times before it converges to a *unique* stable point. Here, we provide a proof sketch illustrating the main ideas behind the theorem and defer the full proof to Appendix E.1.

Proof Sketch. Fix $\theta,\theta' \in \Theta$. Let $f(\varphi) = \mathbb{E}_{Z\sim\mathcal{D}(\theta)}\ell(Z;\varphi)$ and $f'(\varphi) = \mathbb{E}_{Z\sim\mathcal{D}(\theta')}\ell(Z;\varphi)$. By applying standard properties of strong convexity and the fact that $G(\theta)$ is the unique minimizer of $f(\varphi)$, we can derive that,

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

Next, we observe that $(G(\theta) - G(\theta'))^\top\nabla_\theta\ell(z;G(\theta'))$ is $\|G(\theta) - G(\theta')\|_2\beta$-Lipschitz in z. This follows from applying the Cauchy-Schwarz inequality and the fact that the loss is β-jointly smooth. Using the dual formulation of the earth mover's distance (Lemma D.3) and ε-sensitivity of $\mathcal{D}(\cdot)$, as well as the first-order conditions of optimality for convex functions, a short calculation reveals that

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

Claim (a) then follows by combining the previous two inequalities and rearranging. Intuitively, strong convexity forces the iterates to contract after retraining, yet this contraction is offset by the distribution shift induced by changing the underlying classifier. Joint smoothness and ε-sensitivity ensure that this shift is not too large. Part (b) is essentially a consequence of applying the Banach fixed-point theorem to the result of part (a). ■

One intriguing insight from our analysis is that this convergence result is in fact tight; removing any single assumption required for convergence by Theorem 3.5 is enough to construct a counterexample for which RRM diverges.

Proposition 3.6. *Suppose that the distribution map $\mathcal{D}(\cdot)$ is ε-sensitive with $\varepsilon > 0$. RRM can fail to converge at all in any of the following cases, for any choice of $\beta,\gamma > 0$:*

(a) The loss is β-jointly smooth and convex, but not strongly convex.
(b) The loss is γ-strongly convex, but not jointly smooth.
(c) The loss is β-jointly smooth and γ-strongly convex, but $\varepsilon \geqslant \frac{\gamma}{\beta}$.

Proposition 3.6 suggests a fundamental difference between strong and weak convexity in our framing of performative prediction (weak meaning $\gamma = 0$). In supervised learning, using strongly convex losses generally guarantees a

faster rate of optimization, yet asymptotically, the solution achieved with either strongly or weakly convex losses is globally optimal. However, in our framework, strong convexity is in fact *necessary* to guarantee convergence of repeated risk minimization, even for arbitrarily smooth losses and an arbitrarily small sensitivity parameter.

3.3. Repeated Gradient Descent

Theorem 3.5 demonstrates that repeated risk minimization converges to a unique stable point if the sensitivity parameter ε is small enough. However, implementing RRM requires access to an exact optimization oracle. We now relax this requirement and demonstrate how a simple gradient descent algorithm also converges to a unique stable point.

Definition 3.7 (RGD)**.** Repeated gradient descent *(RGD) is the procedure where, starting from an initial model f_{θ_0} we perform the following sequence of updates for $t \geqslant 0$:*

$$\theta_{t+1} = G_{gd}(\theta_t) \stackrel{\text{def}}{=} \Pi_\Theta(\theta_t - \eta \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta_t)} \nabla_\theta \ell(Z;\theta_t)),$$

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

Note that repeated gradient descent only requires the loss ℓ to be differentiable with respect to θ. It does not require taking gradients of the performative risk. Like RRM, we can show that RGD is a contractive mapping for small enough sensitivity parameter ε.

Theorem 3.8. *Suppose that the loss $\ell(z;\theta)$ is β-jointly smooth (A1) and γ-strongly convex (A2). If the distribution map $\mathcal{D}(\cdot)$ is ε-sensitive with $\varepsilon < \frac{\gamma}{(\beta+\gamma)(1+1.5\eta\beta)}$, then RGD with step size $\eta \leqslant \frac{2}{\beta+\gamma}$ satisfies the following:*

(a) $\|G_{gd}(\theta) - G_{gd}(\theta')\|_2$

$$\leqslant \left(1 - \eta\left(\frac{\beta\gamma}{\beta+\gamma} - \varepsilon(1.5\eta\beta^2 + \beta)\right)\right)\|\theta - \theta'\|_2.$$

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

$$\|\theta_t - \theta_{\text{PS}}\|_2 \leqslant \delta \text{ for } t \geqslant \frac{\log\left(\frac{1}{\delta}\|\theta_0 - \theta_{\text{PS}}\|_2\right)}{\eta\left(\frac{\beta\gamma}{\beta+\gamma} - \varepsilon(1.5\eta\beta^2 + \beta)\right)}.$$

The conclusion of Theorem 3.8 is a strict generalization of a classical optimization result which considers a static objective, in which case the rate of contraction is $\left(1 - \eta\frac{\beta\gamma}{\beta+\gamma}\right)$ (see for example Theorem 2.1.15 in Nesterov (2013) or Lemma 3.7 in Hardt et al. (2016b)). Our rate exactly matches this standard result in the case that $\varepsilon = 0$. The proof of Theorem 3.8 can be found in Appendix E.3.

3.4. Finite-Sample Analysis

We now extend our main results regarding the convergence of RRM and RGD to the finite-sample regime. To do so, we leverage the fact that, under mild regularity conditions, the empirical distribution $\mathcal{D}^n$ given by n samples drawn i.i.d. from a true distribution $\mathcal{D}$ is with high probability close to $\mathcal{D}$ in earth mover's distance (Fournier & Guillin, 2015). We begin by defining the finite-sample version of these procedures.

Definition 3.9 (RERM & REGD)**.** *Define* repeated empirical risk minimization *(RERM) to be the procedure where starting from a classifier f_{θ_0} at every iteration $t \geqslant 0$, we collect n_t samples from $\mathcal{D}(\theta_t)$ and perform the update:*

$$\theta_{t+1} = G^{n_t}(\theta_t) \stackrel{\text{def}}{=} \arg\min_\theta \mathop{\mathbb{E}}_{Z\sim\mathcal{D}^{n_t}(\theta_t)} \ell(Z;\theta).$$

Similarly, define repeated empirical gradient descent *(REGD) to be the optimization procedure with update rule:*

$$\theta_{t+1} = G^{n_t}_{gd}(\theta_t) \stackrel{\text{def}}{=} \Pi_\Theta(\theta_t - \eta \mathop{\mathbb{E}}_{Z\sim\mathcal{D}^{n_t}(\theta_t)} \nabla_\theta \ell(Z;\theta_t)).$$

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

The following theorem illustrates that with enough samples collected at every iteration, with high probability both algorithms converge to a small neighborhood around a stable point. Recall that m is the dimension of data samples z.

Theorem 3.10. *Suppose that the loss $\ell(z;\theta)$ is β-jointly smooth (A1) and γ-strongly convex (A2), and that there exist $\alpha > 1, \mu > 0$ such that $\xi_{\alpha,\mu} \stackrel{\text{def}}{=} \int_{\mathbb{R}^m} e^{\mu|x|^\alpha} d\mathcal{D}(\theta)$ is finite $\forall\theta \in \Theta$. Let $\delta \in (0,1)$ be a radius of convergence. Consider running RERM or RGD with $n_t = O\left(\frac{1}{(\varepsilon\delta)^m}\log\left(\frac{t}{p}\right)\right)$ samples at time t.*

(a) If the map $\mathcal{D}(\cdot)$ is ε-sensitive with $\varepsilon < \frac{\gamma}{2\beta}$, then with probability $1-p$, RERM satisfies,

$$\|\theta_t - \theta_{\text{PS}}\|_2 \leqslant \delta, \text{ for all } t \geqslant \frac{\log\left(\frac{1}{\delta}\|\theta_0 - \theta_{\text{PS}}\|_2\right)}{\left(1 - \frac{2\varepsilon\beta}{\gamma}\right)}.$$

(b) If the map $\mathcal{D}(\cdot)$ is ε-sensitive with $\varepsilon < \frac{\gamma}{(\beta+\gamma)(1+1.5\eta\beta)}$, then with probability $1-p$, REGD with satisfies,

$$\|\theta_t - \theta_{\text{PS}}\|_2 \leqslant \delta, \text{ for all } t \geqslant \frac{\log\left(\frac{1}{\delta}\|\theta_0 - \theta_{\text{PS}}\|_2\right)}{\eta\left(\frac{\beta\gamma}{\beta+\gamma} - \varepsilon(3\eta\beta^2 + 2\beta)\right)},$$

for a constant choice of step size $\eta \leqslant \frac{2}{\beta+\gamma}$.

Proof sketch. The basic idea behind these results is the following. While $\|\theta_t - \theta_{\text{PS}}\|_2 > \delta$, the sample size n_t is sufficiently large to ensure a behavior similar to that on a

population level: as in Theorems 3.5 and 3.8, the iterates θ_t contract toward θ_{PS}. This implies that θ_t eventually enters a δ-ball around θ_{PS}, for some large enough t. Once this happens, contrary to population-level results, a contraction is no longer guaranteed due to the noise inherent in observing only finite-sample approximations of $\mathcal{D}(\theta_t)$. Nevertheless, the sample size n_t is sufficiently large to ensure that θ_t cannot escape a δ-ball around θ_{PS} either. ■

4. Relating Optimality and Stability

As we discussed previously, while performative optima are always guaranteed to exist,[1] it is not clear whether performatively stable classifiers exist in all settings. Our algorithmic analysis of repeated risk minimization and repeated gradient descent revealed the existence of unique stable points under the assumption that the objective is strongly convex and smooth. The first result of this section illustrates existence of stable points under weaker assumptions on the loss, in the case where the solution space Θ is constrained. Proofs can be found in Appendix E.

Proposition 4.1. *Let the distribution map $\mathcal{D}(\cdot)$ be ε-sensitive and $\Theta \subset \mathbb{R}^d$ be compact. If the loss $\ell(z;\theta)$ is convex and jointly continuous in (z,θ), then there exists a performatively stable classifier.*

A natural question to consider at this point is whether there are procedures analogous to RRM and RGD for efficiently computing performative optima.

Our analysis suggests that directly minimizing the performative risk is in general a more challenging problem than finding performatively stable points. In particular, we can construct simple examples where the performative risk $\mathrm{PR}(\theta)$ is non-convex, despite strong regularity assumptions on the loss and the distribution map.

Proposition 4.2. *The performative risk $\mathrm{PR}(\theta)$ can be concave in θ, even if the loss $\ell(z;\theta)$ is β-jointly smooth (A1), γ-strongly convex (A2), and the distribution map $\mathcal{D}(\cdot)$ is ε-sensitive with $\varepsilon < \frac{\gamma}{\beta}$.*

However, what we can show is that there are cases where finding performatively stable points is sufficient to guarantee that the resulting classifier has low performative risk. In particular, our next result demonstrates that if the loss function $\ell(z;\theta)$ is Lipschitz in z and γ-strongly convex, then all performatively stable points and performative optima lie in a small neighborhood around each other. Moreover, the theorem holds for cases where performative optima and performatively stable points are not necessarily unique.

Theorem 4.3. *Suppose that the loss $\ell(z;\theta)$ is L_z-Lipschitz in z, γ-strongly convex (A2), and that the distribution map $\mathcal{D}(\cdot)$ is ε-sensitive. Then, for every performatively stable point θ_{PS} and every performative optimum θ_{PO}:*

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

This result shows that in cases where repeated risk minimization converges to a stable point, the resulting classifier approximately minimizes the performative risk.

Moreover, Theorem 4.3 suggests a way of converging close to performative optima *in objective value* even if the loss function is smooth and convex, but not strongly convex. In particular, by adding quadratic regularization to the objective, we can ensure that RRM or RGD converge to a performatively stable point that approximately minimizes the performative risk, see Appendix F.

5. A Case Study in Strategic Classification

Having presented our model for performative prediction, we now proceed to illustrate how these ideas can be applied within the context of strategic classification and discuss some of the implications of our theorems for this setting.

We begin by formally establishing how strategic classification can be cast as a performative prediction problem and illustrate how our framework can be used to derive results regarding the convergence of popular retraining heuristics in strategic classification settings. Afterwards, we further develop the connections between both fields by empirically evaluating the behavior of repeated risk minimization on a dynamic credit scoring task.

5.1. Stackelberg Equilibria are Performative Optima

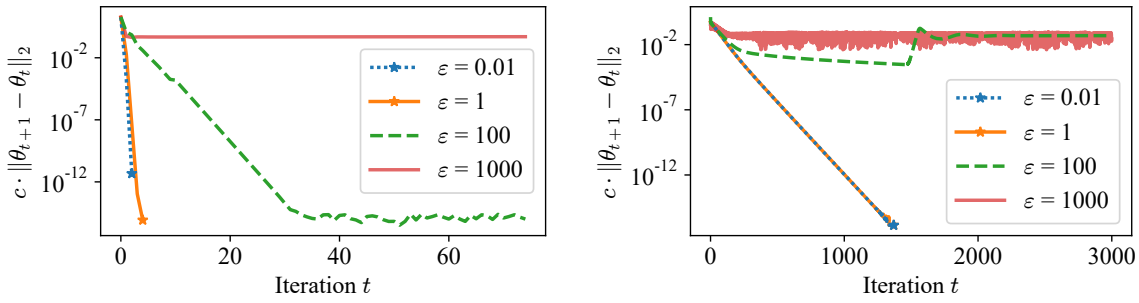
Strategic classification is a two-player game between an institution which deploys a classifier and agents who adapt their features in order to improve their outcomes.

A classic example of this setting is that of a bank which uses a machine learning classifier to predict whether or not a loan applicant is creditworthy. Individual applicants react to the bank's classifier by manipulating their features with the hopes of inducing a favorable classification. This game is said to have a *Stackelberg* structure since agents adapt their features only after the bank has deployed their classifier.

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

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

[1] In particular, they are guaranteed to exist over the extended real line, i.e. we allow $\theta \in (\mathbb{R} \cup \{\pm\infty\})^d$.



(a) Repeated Risk Minimization (RRM)

(b) Repeated Gradient Descent (RGD)

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

We think of $\mathcal{D}$ as a "baseline" distribution over feature-outcome pairs before any classifier deployment, and $\mathcal{D}(\theta)$ denotes the distribution over features and outcomes obtained by strategically manipulating $\mathcal{D}$. As described in previous work (Brückner et al., 2012; Hardt et al., 2016a; Milli et al., 2019), the distribution map $\mathcal{D}(\theta)$ in strategic classification corresponds to the data-generating process given in Figure 2.

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

A strategy commonly adapted in practice as a means of coping with the distribution shift that arises in strategic classification is to repeatedly retrain. This procedure corresponds to the repeated risk minimization procedure introduced in Definition 3.3. Our results describe the first set of sufficient conditions under which repeated retraining overcomes strategic effects.

Corollary 5.1. *Let the institution's loss be L_z- and L_θ-Lipschitz in z and θ respectively, β-jointly smooth (A1), and γ-strongly convex (A2). If the induced distribution map is ε-sensitive, with $\varepsilon < \frac{\gamma}{\beta}$, then RRM converges to a stable classifier θ_{PS} that is $2L_z\varepsilon(L_\theta + L_z\varepsilon)\gamma^{-1}$ close in objective value to the Stackelberg equilibrium.*

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

Figure 2. Distribution map for strategic classification.

5.2. Simulations

We next examine the convergence of repeated risk minimization and repeated gradient descent in a simulated strategic classification setting. We run experiments on a dynamic credit scoring simulator in which an institution classifies the creditworthiness of loan applicants. As motivated previously, agents react to the institution's classifier by manipulating their features to increase the likelihood that they receive a favorable classification.

To run our simulations, we construct a distribution map $\mathcal{D}(\theta)$, as described in Figure 2. For the base distribution $\mathcal{D}$, we use a class-balanced subset of a Kaggle credit scoring dataset (Kaggle, 2012). Features $x \in \mathbb{R}^{m-1}$ correspond to historical information about an individual, such as their monthly income and number of credit lines. Outcomes $y \in \{0, 1\}$ are binary variables which are equal to 1 if the individual defaulted on a loan and 0 otherwise.

The institution makes predictions using a logistic regression classifier. We assume that individuals have linear utilities $u(\theta, x) = -\langle \theta, x \rangle$ and quadratic costs $c(x', x) = \frac{1}{2\varepsilon}\|x' - x\|_2^2$, where ε is a positive constant that regulates the cost incurred by changing features, and hence the sensitivity of the distribution map. Linear utilities indicate that agents wish to minimize their assigned probability of default. We divide the set of features into strategic features $S \subseteq [m-1]$, such as the number of open credit lines, and non-strategic features (e.g., age). Solving the optimization problem described in Figure 2, the best response for an individual corresponds to the following update, $x'_S = x_S - \varepsilon\theta_S$, where $x_S, x'_S, \theta_S \in \mathbb{R}^{|S|}$. As per convention in the literature (Brückner et al., 2012; Hardt et al., 2016a; Milli et al., 2019), individual outcomes y are unaffected by manipulation.

Intuitively, this data-generating process is ε-sensitive since for a given choice of classifiers, f_θ and $f_{\theta'}$, an individual feature vector is shifted to $x_S - \varepsilon\theta_S$ and to $x_S - \varepsilon\theta'_S$, respectively. The distance between these two shifted points is

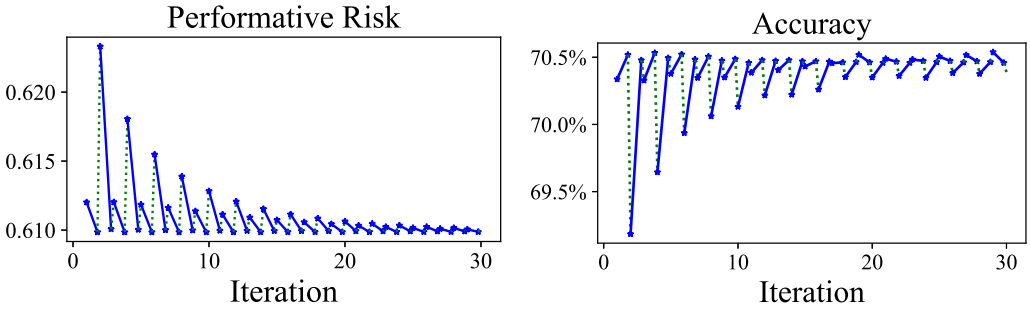


Figure 3. Performative risk (left) and accuracy (right) of the classifier θ_t at different stages of RRM for $\varepsilon = 80$. Solid blue lines indicate the optimization phase and dotted green lines indicate the distribution shift after classifier deployment.

equal to $\varepsilon\|\theta_S - \theta'_S\|_2$. Since the optimal transport distance is bounded by $\varepsilon\|\theta - \theta'\|_2$ for every individual point, it is also bounded by this quantity over the entire distribution. A full proof of this claim is presented in Appendix G.

For our experiments, instead of sampling from $\mathcal{D}(\theta)$, we treat the points in the original dataset as the true distribution. Hence, we can think of all the following procedures as operating on the population level. Furthermore, we add a regularization term to the logistic loss to ensure that the objective is strongly convex. Further details about the experimental setup may be found in Appendix G.

Repeated risk minimization. The first experiment we consider is the convergence of RRM. From our theoretical analysis, we know that RRM is guaranteed to converge at a linear rate to a performatively stable point if the sensitivity parameter ε is smaller than $\frac{\gamma}{\beta}$. In Figure 1 (left), we see that RRM does indeed converge in only a few iterations for small values of ε while it divergences if ε is too large.

The evolution of the performative risk during the RRM optimization is illustrated in Figure 3. We evaluate $PR(\theta)$ at the beginning and at the end of each optimization round and indicate the effect due to distribution shift with a dashed green line. We also verify that the surrogate loss is a good proxy for classification accuracy in the performative setting.

Repeated gradient descent. In the case of RGD, we find similar behavior to that of RRM. While the iterates again converge linearly, they naturally do so at a slower rate than in the exact minimization setting, given that each iteration consists only of a single gradient step. Again, we can see in Figure 1 that the iterates converge for small values of ε and diverge for large values.

6. Discussion and Future Directions

Our work draws attention to the problem of performativity in statistical learning and decision-making. Performative prediction enjoys a clean formal setup that we introduced, drawing on elements from causality and game theory.

Retraining is often considered a nuisance intended to cope with distribution shift. In contrast, our work interprets retraining as the natural equilibrating dynamic for performative prediction. The fixed points of retraining are performative stable points. Moreover, retraining converges to such stable points under natural assumptions, including strong convexity of the loss function. It is interesting to note that (weak) convexity alone is not enough. Performativity thus gives another intriguing perspective on why strong convexity is desirable in supervised learning.

Several interesting questions remain. For example, by letting the step size of repeated gradient descent tend to 0, we see that this procedure converges for $\varepsilon < \frac{\gamma}{\beta+\gamma}$. Exact repeated risk minimization, on the other hand, provably converges for every $\varepsilon < \frac{\gamma}{\beta}$, and we showed this inequality is tight. It would be interesting to understand whether this gap is a fundamental difference between both procedures or an artifact of our analysis.

Lastly, we believe that the tools and ideas from performative prediction can be used to make progress in other subareas of machine learning. For example, in this paper, we have illustrated how reframing strategic classification as a performative prediction problem leads to a new understanding of when retraining overcomes strategic effects. However, we view this example as only scratching the surface of work connecting performative prediction with other fields.

In particular, reinforcement learning can be thought of as a case of performative prediction. In this setting, the choice of policy f_θ, affects the distribution $\mathcal{D}(\theta)$ over $z = \{(s_h, a_h)\}_{h=1}^{\infty}$, the set of visited states, s, and actions, a, in a Markov Decision Process. Building off this connection, we can reinterpret repeated risk minimization as a form of off-policy learning in which an agent first collects a batch of data under a particular policy f_θ, and then finds the optimal policy for that trajectory offline. We believe that some of the ideas developed in the context of performative prediction can shed new light on when these off-policy methods can converge.

Acknowledgements

We wish to acknowledge support from the U.S. National Science Foundation Graduate Research Fellowship Program and the Swiss National Science Foundation Early Postdoc Mobility Fellowship Program.

References

Bartlett, P. L. Learning with a Slowly Changing Distribution. In *Proceedings of the Fifth Annual ACM Conference on Computational Learning Theory (COLT)*, pp. 243–252, 1992.

Bartlett, P. L., Ben-David, S., and Kulkarni, S. R. Learning Changing Concepts by Exploiting the Structure of Change. *Machine Learning*, 41(2):153–174, 2000.

Brückner, M., Kanzow, C., and Scheffer, T. Static Prediction Games for Adversarial Learning Problems. *Journal of Machine Learning Research*, 13(Sep):2617–2654, 2012.

Bubeck, S. Convex Optimization: Algorithms and Complexity. *Foundations and Trends R in Machine Learning*, 8(3-4):231–357, 2015.

Burrell, J., Kahn, Z., Jonas, A., and Griffin, D. When Users Control the Algorithms: Values Expressed in Practices on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, 3:19, 2019.

Dalvi, N., Domingos, P., Sanghai, S., and Verma, D. Adversarial Classification. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 99–108, 2004.

Fournier, N. and Guillin, A. On the Rate of Convergence in Wasserstein Distance of the Empirical Measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. A Survey on Concept Drift Adaptation. *ACM Computing Surveys (CSUR)*, 46(4):1–37, 2014.

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, 2016a.

Hardt, M., Recht, B., and Singer, Y. Train Faster, Generalize Better: Stability of Stochastic Gradient Descent. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 1225–1234, 2016b.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. Fairness Without Demographics in Repeated Loss Minimization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 1929–1938, 2018.

Healy, K. The Performativity of Networks. *European Journal of Sociology/Archives Européennes de Sociologie*, 56 (2):175–205, 2015.

Hu, L. and Chen, Y. A Short-Term Intervention for Long-Term Fairness in the Labor Market. In *Proceedings of the World Wide Web Conference*, pp. 1389–1398, 2018.

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.

Imbens, G. W. and Rubin, D. B. *Causal Inference in Statistics, Social, and Biomedical sciences*. Cambridge University Press, 2015.

Kaggle. Give me Some Credit. https://www.kaggle.com/c/GiveMeSomeCredit/data, 2012.

Khajehnejad, M., Tabibian, B., Schölkopf, B., Singla, A., and Gomez-Rodriguez, M. Optimal Decision Making Under Strategic Behavior. *arXiv preprint arXiv:1905.09239*, 2019.

Kleinberg, J. and Raghavan, M. How Do Classifiers Induce Agents to Invest Effort Strategically? In *Proceedings of the ACM Conference on Economics and Computation (EC)*, pp. 825–844, 2019.

Kuh, A., Petsche, T., and Rivest, R. L. Learning Time-Varying Concepts. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 183–189, 1991.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. Delayed Impact of Fair Machine Learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 3150–3158, 2018.

MacKenzie, D. A., Muniesa, F., and Siu, L. *Do Economists Make Markets?: On the Performativity of Economics*. Princeton University Press, 2007.

Miller, J., Milli, S., and Hardt, M. Strategic Classification is Causal Modeling in Disguise. In *Proceedings of the 37th International Conference on Machine Learning (ICML, to appear)*, 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.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

Pearl, J. *Causality*. Cambridge University Press, 2009.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. ISBN 1107057132.

Villani, C. *Topics in Optimal Transportation*. Number 58. American Mathematical Society, 2003.

Villani, C. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2008.

Webb, G. I., Hyde, R., Cao, H., Nguyen, H. L., and Petitjean, F. Characterizing Concept Drift. *Data Mining and Knowledge Discovery*, 30(4):964–994, 2016.

Žliobaitė, I. Learning under Concept Drift: an Overview. *arXiv preprint arXiv:1010.4784*, 2010.

Žliobaitė, I., Pechenizkiy, M., and Gama, J. An Overview of Concept Drift Applications. In *Big Data Analysis: New Algorithms for a New Society*, pp. 91–114. Springer, 2016.