Performative Prediction in a Stateful World

Gavin Brown[*]
Boston University
grbrown@bu.edu

Shlomi Hod[*]
Boston University
shlomi@bu.edu

Iden Kalemaj[*]
Boston University
ikalemaj@bu.edu

Abstract

Deployed supervised machine learning models make predictions that interact with and influence the world. This phenomenon is called *performative prediction* by Perdomo et al. (2020), who investigated it in a stateless setting. We generalize their results to the case where the response of the population to the deployed classifier depends both on the classifier and the previous distribution of the population. We also demonstrate such a setting empirically, for the scenario of strategic manipulation.

1 Introduction

Supervised learning is widely used to train classifiers that aid institutions in decision-making: will a loan applicant default? Will a user respond well to certain recommendations? Will a candidate perform well in this job? Several studies and examples suggest that such predictions can influence the behavior of the target population that they try to predict [2, 5, 7]. Loan applicants strategically manipulate credit card usage to appear more creditworthy, job applicants tailor their resumes to resume-parsing algorithms, and user preferences on a platform shift as they interact with recommended items. It is an ongoing challenge to understand the influence of such predictions and design tools so as to control that influence. Specifically, this influence can manifest as a distribution shift in the target population, causing a loss in the prediction accuracy of the classifier with respect to the new distribution and creating the need for a new classifier.

In their recent paper "Performative Prediction," Perdomo, Zrnic, Mendler-Dünner, and Hardt [6] term such predictions *performative*. They establish a theoretical framework for analyzing performativity in supervised learning and propose *repeated risk minimization* as a strategy that institutions can apply in hopes of converging to an equilibrium. They propose that the equilibrium should be a classifier that is optimal for the distribution it induces. Perdomo et al. model the response of the target population via a deterministic function of the published classifier θ. The distribution induced by a classifier θ is unaffected by previously-deployed classifiers. However, in practice the environment may depend heavily on the history of classifiers deployed by an institution.

Consider the following example: individuals applying for loans manipulate their features to receive favorable results from a published classifier. The cost of this manipulation depends not only on the classifier, but also on the previous feature state of the individual. As the bank updates their criteria for creditworthiness, new features may become important. Without recording the previous state, we cannot model the strategic behavior of individuals in the current state.

In addition, repeated classification can induce structural changes and affect individuals over generations. For example, learning to code is increasingly easy with courses that offer pay-when-employed programs, and efforts are underway to make programming qualifications more accessible for minority groups; such qualifications were harder to obtain ten years ago. These changes are not just the result of current policies, but also of the long history of classification. Finally, when a new classifier is published, different groups in the target population acquire information and adapt their behavior

[*]Equal contribution.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

at different rates, so that repeated application of the same classification model may still result in a distribution shift.

1.1 Our framework

We cast the phenomenon of performativity in repeated decision-making as an online learning game [8]. This view differs slightly from that of Perdomo et al., and we find it captures well the dynamics of performativity and eases our discussion. At round t, the learner/institution chooses a classifier θ_t to publish. In response, the adversary/environment picks a distribution d_t over labeled samples. The institution then suffers loss $\mathbb{E}_{Z\sim d_t}[\ell(Z;\theta_t)]$ for some fixed loss function ℓ. We ignore finite-sample issues and assume that the institution observes the distribution directly.

Standard online learning assumes that the adversary may be malicious and pick whichever distribution causes the greatest loss. To model state and performativity, we propose a weaker adversary that responds according to a *transition map* $\texttt{Tr}(;)$, mapping classifier-distribution pairs to distributions. The transition map is fixed but a priori unknown to the institution. If the institution plays θ and the previous distribution played by the adversary was d, the adversary must respond with

$$d' = \texttt{Tr}(d;\theta).$$

We denote by $\theta_1, \theta_2, \dots$ the classifiers played by the institution, and by $d_1, d_2, \dots$ the distributions played by the adversary.

Our framework generalizes the framework of Perdomo et al., in which the transition map is independent of the distribution d and depends only on the current classifier θ. Our key conceptual contribution is incorporating the current state/distribution of the target population into their performative response via the two arguments of the transition map $\texttt{Tr}(;)$. We call our framework *stateful*, since it preserves information about the state of the world and the history of classifiers played by the institution. This is in contrast to the *stateless* framework of Perdomo et al.

A particular phenomenon captured by our framework is that of individuals acting strategically but with outdated information. We formalize this behavior in the two examples below, which may be of independent interest for the study of performativity and group fairness. We also demonstrate these two examples empirically in Section 3.

Example 1 (Geometric decay response)**.** Assume there is a deterministic "strategic response function" $\mathcal{D}(\theta)$ unknown to the learner. The adversary plays a mixture over past responses. For $\delta \in [0,1]$, define

$$\texttt{Tr}(d_{t-1};\theta_t) = (1-\delta)d_{t-1} + \delta \cdot \mathcal{D}(\theta_t).$$

The mixture coefficients in the current distribution decay geometrically across older responses. When $\delta = 1$ this is the setting of Perdomo et al.

Example 2 (k Groups respond slowly)**.** Assume again an unknown $\mathcal{D}(\theta)$. Suppose there are k groups, where group $j \in [k]$ responds to the classifier from j rounds ago. For distribution d, let $d^{|j}$ be the distribution conditioned on being in group j. Then in response to θ_t, the individuals update:

$$d_t^{|1} = \mathcal{D}(\theta_t) \qquad \text{and} \qquad \forall j > 1, d_t^{|j} = d_{t-1}^{|j+1}.$$

This corresponds to a transition function between mixtures of distributions. It provides a simple model of distinct groups who receive information at different rates. When $k = 1$ this is the setting of Perdomo et al.

1.2 Our Results

Our goal is to devise a strategy for the institution which converges towards an approximately-optimal distribution-classifier pair. Perdomo et al. propose the strategy of *repeated risk minimization* (RRM) where, at every round, the institution chooses the classifier that minimizes loss on the last distribution played by the adversary:

$$\theta_{t+1} = \operatorname*{argmin}_{\theta} \mathop{\mathbb{E}}_{Z\sim d_t} \ell(Z;\theta).$$

It is a natural strategy and akin to many retraining heuristics used in practice to deal with different kinds of distribution shifts. Perdomo et al. analyze RRM in the stateless framework and show that,

under convexity and Lipschitz assumptions, it will converge to a near-optimal classifier. It is unclear if RRM converges when the history of previous classifiers can influence the distribution. For instance, it might demonstrate "thrashing" behavior. We illustrate this in Example 3.

Nevertheless, we are able to provide conditions under which RRM will converge. We add assumptions that control the extent of performative response allowed by the transition function. Similarly to Perdomo et al., we impose a Lipschitz requirement on the transition map to ensure that small changes in the distribution or the classifier only yield small changes in the updated distribution. When the Lipschitz constant is strictly less than 1, the transition map is contractive under repeated application of any classifier, and so the distribution converges to a fixed point. The concept of a fixed point distribution for every classifier is a key aspect of our framework and results. Intuitively, this models behavior where the environment will eventually settle on a response to the institution's classifier.

Our first result shows sufficient conditions for convergence of repeated risk minimization. Convergence is to an equilibrium point: a fixed point distribution and a classifier that achieves minimum loss on this distribution.

Theorem 1 (Informal)**.** *If the loss function $\ell(z;\theta)$ is smooth and strongly convex and the transition map $\texttt{Tr}(d;\theta)$ is Lipschitz in both arguments, then repeated risk minimization converges to an equilibrium distribution-classifier pair.*

Next, we formalize the notion of an *optimal* strategy for a institution faced with the phenomenon of performativity. We show that repeated risk minimization also provides a means to approximate such optimal strategies.

Theorem 2 (Informal)**.** *If the loss function $\ell(z;\theta)$ is Lipschitz and strongly convex and the transition map $\texttt{Tr}(d;\theta)$ is Lipschitz in both arguments, all equilibrium points and optimal points lie within a small distance of each other.*

Theorem 1 and Theorem 2, which we state formally in Section 2, generalize results of Perdomo et al. for the stateless framework to our stateful framework. We include proofs in Appendix A.

2 Framework and main results

In this section we formally state our main results and the relevant definitions. We parameterize machine learning models by real-valued vectors $\theta \in \Theta$, where the classification space Θ is a closed, convex subset of $\mathbb{R}^d$. In round t, the institution chooses a classifier θ_t. The adversary responds with a distribution d_t over instances $Z = (X, Y)$ of feature-label pairs. Let $\Delta(\mathcal{Z})$ denote the space of distributions. The adversary is restricted to pick its distribution according to a deterministic transition map:

$$\texttt{Tr} : \Theta \times \Delta(\mathcal{Z}) \to \Delta(\mathcal{Z}),$$

so that $\texttt{Tr}(d_{t-1}, \theta_t) = d_t$. We assume that an initial distribution d_0 is publicly known. From this online game view, the work of Perdomo et al. assumes a myopic adversary that *only* sees the institution's most recent play. We remove that condition.

Algorithm 1 Performative prediction with state

1: initial distribution $d_0 \in \Delta(\mathcal{Z})$ ▷ Publicly known
2: **for** t = 1,2,... **do**
3: Institution plays $\theta_t \in \Theta$.
4: Adversary computes $d_t = \texttt{Tr}(d_{t-1}; \theta_t)$. ▷ $\texttt{Tr}(;)$ function unknown to institution
5: Institution observes d_t, suffers loss $\mathbb{E}_{Z\sim d_t}[\ell(Z;\theta_t)]$.

2.1 Repeated risk minimization and stable points

Perdomo et al. propose the following strategy for the institution: at each round, play the classifier that minimizes loss on the previous distribution. We investigate the same strategy in our stateful framework.

Definition 1 (Repeated risk minimization (RRM)). *Denote by $G(d)$ the updated classifier[2]:*

$$G(d) := \operatorname*{argmin}_{\theta'} \mathop{\mathbb{E}}_{Z\sim d} \ell(Z;\theta').$$

Following the notation of Algorithm 1, at round t, the institution updates its classifier to $\theta_t = G(d_{t-1})$.

We consider two sufficient conditions for the convergence of RRM in objective value: approaching a *fixed point distribution* and approaching a classifier that is optimal for this distribution.

Definition 2 (Fixed point distribution). *A distribution d_θ is a fixed point for θ if $\texttt{Tr}(d_\theta, \theta) = d_\theta$.*

Definition 3 (Stable points). *A distribution-classifier pair $(d_{\mathrm{PS}}, \theta_{\mathrm{PS}})$ is a performatively stable point if the following hold:*

1. $\texttt{Tr}(d_{\mathrm{PS}}, \theta_{\mathrm{PS}}) = d_{\mathrm{PS}}$*, i.e. d_{PS} is a fixed point distribution for θ_{PS}.*
2. $\theta_{\mathrm{PS}} = G(d_{\mathrm{PS}})$*, i.e. θ_{PS} minimizes the loss on d_{PS}.*

Once the game approaches the distribution d_{PS}, the institution can repeatedly play θ_{PS} with no need for retraining, while incurring the lowest possible loss on the distribution d_{PS}. It is not obvious however that such stable points exist for every setting. Nevertheless, we shows sufficient conditions on the loss and transition function for RRM to converge to a stable point.

Definition 4 (ε-joint sensitivity). *The transition map $\texttt{Tr}(;)$ is ε-jointly sensitive if, for all $\theta, \theta' \in \Theta$ and $d, d' \in \Delta(\mathcal{Z})$,*

$$\mathcal{W}_1(\texttt{Tr}(d;\theta), \texttt{Tr}(d';\theta')) \leq \varepsilon\|\theta - \theta'\|_2 + \varepsilon\mathcal{W}_1(d, d'),$$

where $\mathcal{W}_1$ denotes the Wasserstein-1 distance between distributions.

Definition 5 (Strong convexity). *A loss function $\ell(z;\theta)$ is γ-strongly convex if, for all $\theta, \theta' \in \Theta$ and $z \in \mathcal{Z}$,*

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

Definition 6 (Smoothness). *A loss function $\ell(z;\theta)$ is β-jointly smooth if the gradient with respect to θ is β-Lipschitz in θ and z:*

$$\|\nabla_\theta \ell(z;\theta) - \nabla_\theta \ell(z;\theta')\|_2 \leq \beta\|\theta - \theta'\|_2, \quad \|\nabla_\theta \ell(z;\theta) - \nabla_\theta \ell(z';\theta)\|_2 \leq \beta\|z - z'\|_2,$$

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

Finally, our theorems are clearer with additional notation wrapping the institution's and adversary's actions into one step.

Definition 7 (RRM map). *Define the RRM map $f\colon \Delta(\mathcal{Z}) \times \Theta \to \Delta(\mathcal{Z}) \times \Theta$ as:*

$$f(d, \theta) = (\texttt{Tr}(d, \theta), G(\texttt{Tr}(d, \theta))).$$

In our game, $f(d_{t-1}, \theta_t) = (d_t, G(d_t)) = (d_t, \theta_{t+1})$.

We endow the space $\Delta(\mathcal{Z}) \times \Theta$ with the product metric dist, so that:

$$\mathrm{dist}((d, \theta), (d', \theta')) = \mathcal{W}_1(d, d') + \|\theta - \theta'\|_2.$$

The next example shows that, without the right interplay of the above parameters, there are settings for which RRM may not converge.

Example 3 (RRM may not converge). Suppose that the loss function is the squared loss $\ell(z;\theta) = (y-\theta)^2$ for $\theta \in [1, \infty)$. The loss function is β-jointly smooth and γ-strongly convex, with $\beta = \gamma = 2$. Consider the transition map $\texttt{Tr}(d;\theta) = 1 + \varepsilon d + \varepsilon\theta$, which operates on point mass distributions $d \in [1, \infty)$ of the outcome Y. Clearly, the transition function $\texttt{Tr}$ is ε-jointly sensitive. Finally, let some $d_0 \in [1, \infty)$ be the starting distribution of the game.

When the institution uses RRM starting from d_0, we get that:

$$\theta_{t+1} = G(d_t) = \operatorname*{argmin}_{\theta} \mathop{\mathbb{E}}_{Z\sim d_t} \ell(Z;\theta) = d_t,$$
$$d_{t+1} = \texttt{Tr}(d_t;\theta_{t+1}) = 1 + \varepsilon d_t + \varepsilon\theta_{t+1}.$$

[2]For the scenarios we consider, the set $\mathrm{argmin}_{\theta'}$ will be non-empty. When the set has more than one element, we choose a value for $G(d)$ from the set arbitrarily.

Consequently, $\theta_{t+2} = d_{t+1} = 1 + \varepsilon d_t + \varepsilon\theta_{t+1} = 1 + 2\varepsilon\theta_{t+1}$.

The distance between two successive θs is $|\theta_{t+2} - \theta_{t+1}| = |(1+2\varepsilon\theta_{t+1}) - (1+2\varepsilon\theta_t|) = 2\varepsilon|\theta_{t+1} - \theta_t|$.

If we only require $\varepsilon < \frac{\gamma}{\beta} = 1$, then whenever $\varepsilon > \frac{1}{2}$, the sequence of θs produced by RRM fails to converge. For $\frac{\varepsilon}{1-\varepsilon} < \frac{\beta}{\gamma}$, i.e. $\varepsilon < \frac{1}{2}$, the sequence converges.

Our main result is sufficient conditions for the convergence of RRM in the stateful framework.

Theorem 3. *Suppose the transition map* $\texttt{Tr}(;)$ *is* ε*-jointly sensitive with* $\varepsilon < 1$*, the loss function* $\ell(z;\theta)$ *is* β*-jointly smooth and* γ*-strongly convex. Then for the RRM map* f *it holds that:*

(a) $\operatorname{dist}(f(d,\theta), f(d',\theta')) \leq \frac{\varepsilon}{1-\varepsilon}\frac{\beta}{\gamma}\|\theta - \theta'\|_2 + \frac{\varepsilon}{1-\varepsilon}\frac{\beta}{\gamma}\mathcal{W}_1(d,d')$.

(b) In particular, if $\frac{\varepsilon}{1-\varepsilon} < \frac{\gamma}{\beta}$*, then* f *has a unique fixed point which is a stable point with respect to* $\texttt{Tr}(;)$*. RRM will converge to this stable point at a linear rate.*

Our conditions for convergence of RRM are similar to, but stricter, than those of the setting of Perdomo et al. In particular, their results only require that $\varepsilon < \frac{\gamma}{\beta}$. Ex. 3 shows that our stricter requirement is necessary.

2.2 Performative optimality

Theorem 3 guarantees that RRM converges to an equilibrium, but this stable point might not be optimal in a more global sense. In fact, it is not obvious how to define optimal strategies within our framework. Since the sequence of distributions played by the adversary depends on the initial distribution d_0, the best possible strategy for the learner might depend on d_0. This is further complicated by the fact that repeatedly playing the same fixed classifier might result in a distribution shift. Therefore, we restrict our attention to scenarios where the transition map $\texttt{Tr}(;)$ is ε-jointly sensitive with $\varepsilon < 1$. In that case, repeated application of the same classifier θ is guaranteed to converge to a fixed point distribution (Definition 2).

Claim 1. *If the transition map* $\texttt{Tr}(;)$ *is* ε*-jointly sensitive with* $\varepsilon < 1$*, then for each* $\theta \in \Theta$*, there exists a unique fixed point distribution* d_θ*. For any starting distribution* d_0*, iterated application of the same classifier* θ*, will result in a sequence of distributions that converges to* d_θ *at a linear rate.*

Claim 1 follows immediately from Banach's fixed point theorem.

Our definition of the optimal strategy considers the "long-run" loss of a fixed classifier. Assume the institution plays the same fixed classifier θ for all rounds of the game. We measure the long-run loss of θ as the loss on its corresponding fixed point distribution d_θ. The optimal θ is the one which minimizes its long-run loss.

Definition 8 (Performative optimality)**.** *The long-run loss of a classifier* θ *is the loss* $\mathop{\mathbb{E}}_{Z\sim d_\theta} \ell(Z;\theta)$*, where* d_θ *denotes the unique fixed point distribution for the classifier* θ*. A classifier* θ_{PO} *is performatively optimal if achieves the minimum long-run loss amongst all classifiers in* Θ*.*

If an institution had prior knowledge of the transition map, a reasonable strategy would be to play the fixed classifier θ_{PO} for all rounds of classification. We note that θ_{PO} is undefined when no classifier has a fixed point distribution. However, it is guaranteed to exist when $\texttt{Tr}(;)$ is ε-jointly sensitive with $\varepsilon < 1$.

Our definitions of stable and optimal points generalize those of Perdomo et al. for the stateless framework. As pointed out in Perdomo et al., for a given setting, the optimal classifier does not necessarily coincide with a stable point. Our next result shows that RRM approximately approaches optimal points.

Theorem 4. *Suppose that the loss* $\ell(z;\theta)$ *is* L_z*-Lipschitz,* γ*-strongly convex, and that the transition map is* ε*-jointly sensitive with* $\varepsilon < 1$*. Then for every stable point* θ_{PS} *and optimal point* θ_{PO}*:*

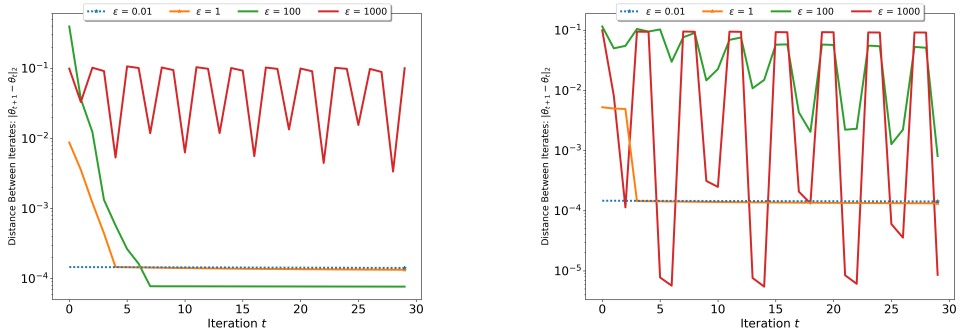
$$\|\theta_{\mathrm{PO}} - \theta_{\mathrm{PS}}\|_2 \leq \frac{2L_z\varepsilon}{\gamma(1-\varepsilon)}.$$

3 Simulation

Strategic classification studies the behavior of individuals who wish to achieve a more preferable outcome from a classifier by manipulating their attributes without changing their true label [3]. It is one instantiation of performative prediction. We adapt a simulation of loan applications in Perdomo et al. and enrich it with state. We demonstrate the convergence of RRM for the scenarios of Geometric decay response (Ex. 1) and k Groups respond slowly (Ex. 2) in a credit score setting[3].

The baseline distribution of the population is determined by Kaggle's *GiveMeSomeCredit* dataset [4]. An individual's strategic response is based on cost and utility functions which take into account the published classifier and the feature state of the individual. The parameter ε controls the strength of the strategic response; larger values allow greater manipulation. Refer to Appendix B for a detailed description of the mechanics.

Figure 1 shows the game dynamics when an institution uses RRM in both scenarios. The two games converge to equilibrium much faster for lower values of ε. Interestingly, a periodic behavior surfaces for large values of ε, which is not the case in the stateless simulation of Perdomo et al. This behavior shows that RRM is not well-suited to settings where state plays an important role and the population reacts in non-smooth ways. For such settings, other types of learning strategies might be more adequate.



(a) Geometric decay response with $\delta = 0.7$.

(b) Three groups respond slowly.

Figure 1: Convergence of repeated risk minimization for varying values of ε. The horizontal axis shows the number of iterations and the vertical axis shows the distance between successive θs.

4 Discussion

This work points out the important role that the long history of institutions making predictions on individuals plays in shaping the behavior of individuals. The addition of state to performativity opens up a new venue for discussing the social impact of machine learning based predictions. Examples include the structural changes that enable individuals to succeed under such modes of classification and the disparate impact of predictions on groups over time.

It remains open whether other algorithms studied in online learning can yield successful outcomes for the phenomenon of performativity. This work focused on the goal of convergence, but if convergence is not a priority for the institution, then it is interesting to study which measures of the learner's success best apply to the setting of performativity. Regret is widely studied in online learning, but lacks a clear interpretation in settings where the adversary can adapt to the player. Empirically, studying other applications where performativity arises could provide important theoretical insight into this phenomenon.

[3] Simulation code: `https://github.com/shlomihod/performative-prediction-stateful-world`

Broader impact

Our work assumes a simple and abstract model of repeated decision-making. While the study of performativity in prediction may have wide social effects in general, we do not believe this paper will have direct ethical or social consequences.

Acknowledgments

The authors would like to thank Adam D. Smith for fruitful conversations and guidance, and Ran Canetti for helpful comments.

References

[1] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3-4):231–357, 2015.

[2] Adriana Camacho and Emily Conover. Manipulation of social program eligibility. *American Economic Journal: Economic Policy*, 3(2):41–65, 2011.

[3] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

[4] Kaggle. Give me some credit dataset. `https://www.kaggle.com/c/GiveMeSomeCredit`, 2011.

[5] Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

[6] Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. *CoRR*, abs/2002.06673, 2020.

[7] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 131–141, 2020.

[8] Shai Shalev-Shwartz. Online learning and online convex optimization. *Found. Trends Mach. Learn.*, 4(2):107–194, 2012.

A Proofs of main theorems

We first state two key lemmas used in the proofs of Theorem 3 and Theorem 4. Proofs are omitted. The proof of the first lemma uses the Kantorovich-Rubinstein duality of $\mathcal{W}_1$.

Lemma 1. *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be β-Lipschitz, and let X, X' be random variables. Then*

$$\|\mathbb{E}[f(X)] - \mathbb{E}[f(X')]\|_2 \leq \beta \cdot \mathcal{W}_1(X, X').$$

Lemma 2 ([1])**.** *If g is convex and Ω is a closed convex set on which g is differentiable, and*

$$x_* \in \operatorname*{argmin}_{x \in \Omega} g(x),$$

then:

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

A.1 Proof of Theorem 3

Proof of Theorem 3. Note that if part (a) holds, then part (b) follows from the fact that the map f is contractive with contraction coefficient $\frac{\varepsilon}{1-\varepsilon}\frac{\beta}{\gamma} < 1$. By the Banach fixed point theorem, f has a unique fixed point. Suppose (d^*, θ^*) is the fixed point of f, so that $f(d^*, \theta^*) = (d^*, \theta^*)$. This point is also a stable point for it satisfies: $d^* = \mathtt{Tr}(d^*, \theta^*)$ and $\theta^* = G(\mathtt{Tr}(d^*, \theta^*)) = G(d^*)$.

We now show part (a). First we will simplify notation and let $G(d, \theta) := G(\mathtt{Tr}(d, \theta))$. By definition of f:

$$\begin{aligned}\operatorname{dist}(f(d,\theta), f(d',\theta')) &= \operatorname{dist}((\mathtt{Tr}(d,\theta), G(d,\theta)), (\mathtt{Tr}(d',\theta'), G(d',\theta'))) \\ &= \mathcal{W}_1(\mathtt{Tr}(d,\theta), \mathtt{Tr}(d',\theta')) + \|G(d,\theta) - G(d',\theta')\|_2.\end{aligned}$$

The ε-joint sensitivity of the transition map yields:

$$\mathcal{W}_1(\mathtt{Tr}(d,\theta), \mathtt{Tr}(d',\theta')) \leq \varepsilon \mathcal{W}_1(d, d') + \varepsilon\|\theta - \theta'\|_2. \tag{1}$$

We will show that:

$$\|G(d,\theta) - G(d',\theta')\|_2 \leq \varepsilon\frac{\beta}{\gamma}\mathcal{W}_1(d, d') + \varepsilon\frac{\beta}{\gamma}\|\theta - \theta'\|_2. \tag{2}$$

Combining equations (1) and (2) will conclude the proof.

We obtain (2) from a slight modification of the proof of Theorem 3.5 of Perdomo et al. By Lemma 2 we obtain:

$$\begin{aligned}(G(d',\theta') - G(d,\theta))^\top \mathop{\mathbb{E}}_{Z \sim \mathtt{Tr}(d,\theta)} \nabla_\theta \ell(Z; G(d,\theta)) \geq 0, \\ (G(d,\theta) - G(d',\theta'))^\top \mathop{\mathbb{E}}_{Z \sim \mathtt{Tr}(d',\theta')} \nabla_\theta \ell(Z; G(d',\theta')) \geq 0.\end{aligned}$$

Combine these two inequalities:

$$(G(d,\theta) - G(d',\theta'))^\top \big(\mathop{\mathbb{E}}_{Z \sim \mathtt{Tr}(d,\theta)} \nabla_\theta \ell(Z; G(d,\theta)) - \mathop{\mathbb{E}}_{Z \sim \mathtt{Tr}(d',\theta')} \nabla_\theta \ell(Z; G(d',\theta')) \big) \leq 0.$$

Then add and subtract a term to obtain:

$$\begin{aligned}&(G(d,\theta) - G(d',\theta'))^\top \big(\mathop{\mathbb{E}}_{Z \sim \mathtt{Tr}(d,\theta)} \nabla_\theta \ell(Z; G(d,\theta)) - \mathop{\mathbb{E}}_{Z \sim \mathtt{Tr}(d',\theta')} \nabla_\theta \ell(Z; G(d,\theta)) \big) \\ +&(G(d,\theta) - G(d',\theta'))^\top \big(\mathop{\mathbb{E}}_{Z \sim \mathtt{Tr}(d',\theta')} \nabla_\theta \ell(Z; G(d,\theta)) - \mathop{\mathbb{E}}_{Z \sim \mathtt{Tr}(d',\theta')} \nabla_\theta \ell(Z; G(d',\theta')) \big) \leq 0.\end{aligned} \tag{3}$$

For the first term of the sum, note that the function $(G(d,\theta) - G(d',\theta'))^\top \nabla_\theta \ell(z; G(d,\theta))$ is $(\|G(d,\theta) - G(d',\theta')\|_2 \beta)$-Lipschitz in z. This follows from the Cauchy-Schwarz inequality and the fact that the loss is β-smooth. We can thus bound the first term of (3) as:

$$\begin{aligned}&(G(d,\theta) - G(d',\theta'))^\top \big(\mathop{\mathbb{E}}_{Z \sim \mathtt{Tr}(d,\theta)} \nabla_\theta \ell(Z; G(d,\theta)) - \mathop{\mathbb{E}}_{Z \sim \mathtt{Tr}(d',\theta')} \nabla_\theta \ell(Z; G(d,\theta)) \big) \\ \geq& -\beta\|G(d,\theta) - G(d',\theta')\|_2(\varepsilon \cdot \mathcal{W}_1(d, d') + \varepsilon\|\theta - \theta'\|_2),\end{aligned}$$

where we use the property that $\texttt{Tr}(\cdot;\cdot)$ is ε-jointly sensitive together with Lemma 1.

For the second term of the sum in (3) we invoke the γ-convexity of the loss:

$$\begin{aligned}&(G(d,\theta)-G(d',\theta'))^\top\big(\mathop{\mathbb{E}}_{Z\sim\texttt{Tr}(d',\theta')}\nabla_\theta\ell(Z;G(d,\theta))-\mathop{\mathbb{E}}_{Z\sim\texttt{Tr}(d',\theta')}\nabla_\theta\ell(Z;G(d',\theta'))\big)\\ \geq\quad&\gamma\|G(d,\theta)-G(d',\theta')\|_2^2.\end{aligned}$$

Replace the last two inequalities into (3):

$$0\geq-\beta\|G(d,\theta)-G(d',\theta')\|_2(\varepsilon\cdot\mathcal{W}_1(d,d')+\varepsilon\|\theta-\theta'\|_2)+\gamma\|G(d,\theta)-G(d',\theta')\|_2^2.$$

After canceling out $\|G(d,\theta)-G(d',\theta')\|_2$ and rearranging we conclude:

$$\|G(d,\theta)-G(d',\theta')\|_2\leq\varepsilon\frac{\beta}{\gamma}\|\theta-\theta'\|_2+\varepsilon\frac{\beta}{\gamma}\mathcal{W}_1(d,d').$$

□

A.2 Proof of Theorem 4

The following lemma is used in the proof of Theorem 4.

Lemma 3. *Suppose the map* $\texttt{Tr}(;)$ *is* ε*-jointly sensitive with* $\varepsilon<1$*. Then for any* θ_1,θ_2 *and their corresponding fixed point distributions* d_1,d_2 *it holds that:*

$$\mathcal{W}_1(d_1,d_2)\leq\frac{\varepsilon}{1-\varepsilon}\|\theta_1-\theta_2\|_2.$$

Proof. Note that $\mathcal{W}_1(d_1,d_2)=\mathcal{W}_1(\texttt{Tr}(d_1,\theta_1),\texttt{Tr}(d_2,\theta_2))$, from the definition of fixed point distributions. By the ε-joint sensitivity of the transition map:

$$\mathcal{W}_1(d_1,d_2)\leq\varepsilon\mathcal{W}_1(d_1,d_2)+\varepsilon\|\theta_1-\theta_2\|_2.$$

The statement follows from the equation above. □

Proof of Theorem 4. Let θ_{PO} be an optimal classifier and let d_{PO} be its corresponding fixed point distribution. Let θ_{PS} be a stable point, with corresponding fixed point distribution d_{PS}. By the definitions of optimality and stability:

$$\mathop{\mathbb{E}}_{Z\sim d_{\mathrm{PO}}}\ell(Z;\theta_{\mathrm{PO}})\leq\mathop{\mathbb{E}}_{Z\sim d_{\mathrm{PS}}}\ell(Z;\theta_{\mathrm{PS}})\leq\mathop{\mathbb{E}}_{Z\sim d_{\mathrm{PS}}}\ell(Z;\theta_{\mathrm{PO}})\,.$$

We first show that:

$$\mathop{\mathbb{E}}_{Z\sim d_{\mathrm{PS}}}\ell(Z;\theta_{\mathrm{PO}})-\mathop{\mathbb{E}}_{Z\sim d_{\mathrm{PS}}}\ell(Z;\theta_{\mathrm{PS}})\geq\frac{\gamma}{2}\|\theta_{\mathrm{PO}}-\theta_{\mathrm{PS}}\|_2^2\,. \tag{4}$$

By the strong convexity of the loss function we know that for all z:

$$\ell(z;\theta_{\mathrm{PO}})\geq\ell(z;\theta_{\mathrm{PS}})+\nabla_\theta\ell(z;\theta_{\mathrm{PS}})^\top(\theta_{\mathrm{PO}}-\theta_{\mathrm{PS}})+\frac{\gamma}{2}\|\theta_{\mathrm{PO}}-\theta_{\mathrm{PS}}\|_2^2\,.$$

As a result:

$$\mathop{\mathbb{E}}_{Z\sim d_{\mathrm{PS}}}\big[\ell(Z;\theta_{\mathrm{PO}})-\ell(Z;\theta_{\mathrm{PS}})\big]\geq\mathop{\mathbb{E}}_{Z\sim d_{\mathrm{PS}}}\big[\nabla_\theta\ell(z;\theta_{\mathrm{PS}})^\top(\theta_{\mathrm{PO}}-\theta_{\mathrm{PS}})\big]+\frac{\gamma}{2}\|\theta_{\mathrm{PO}}-\theta_{\mathrm{PS}}\|_2^2\,.$$

Since θ_{PS} minimizes the value of ℓ over the distribution d_{PS}, Lemma 2 implies:

$$\mathop{\mathbb{E}}_{Z\sim d_{\mathrm{PS}}}\big[\nabla_\theta\ell(Z;\theta_{\mathrm{PS}})^\top(\theta_{\mathrm{PO}}-\theta_{\mathrm{PS}})\big]\geq0.$$

Therefore, (4) holds. On the other hand, since the loss is L_z-Lipschitz in z, and by Lemma 3:

$$\mathop{\mathbb{E}}_{Z\sim d_{\mathrm{PS}}}\ell(Z;\theta_{\mathrm{PO}})-\mathop{\mathbb{E}}_{Z\sim d_{\mathrm{PO}}}\ell(Z;\theta_{\mathrm{PO}})\leq L_z\mathcal{W}_1(d_{\mathrm{PS}},d_{\mathrm{PO}})\leq\frac{L_z\varepsilon}{1-\varepsilon}\|\theta_{\mathrm{PO}}-\theta_{\mathrm{PS}}\|_2\,.$$

If $\frac{\varepsilon}{1-\varepsilon}<\frac{\gamma\|\theta_{\mathrm{PO}}-\theta_{\mathrm{PS}}\|_2}{2L_z}$ then $\frac{L_z\varepsilon}{1-\varepsilon}\|\theta_{\mathrm{PO}}-\theta_{\mathrm{PS}}\|_2\leq\frac{\gamma}{2}\|\theta_{\mathrm{PO}}-\theta_{\mathrm{PS}}\|_2^2$. This is a contradiction since we argued that:

$$\mathop{\mathbb{E}}_{Z\sim d_{\mathrm{PS}}}\ell(Z;\theta_{\mathrm{PO}})-\mathop{\mathbb{E}}_{Z\sim d_{\mathrm{PO}}}\ell(Z;\theta_{\mathrm{PO}})\geq\mathop{\mathbb{E}}_{Z\sim d_{\mathrm{PS}}}\ell(Z;\theta_{\mathrm{PO}})-\mathop{\mathbb{E}}_{Z\sim d_{\mathrm{PS}}}\ell(Z;\theta_{\mathrm{PS}}).$$

□

B Simulation details

We simulate a credit score system using the dataset *GiveMeSomeCredit* [4] from Kaggle. Before giving a loan to an applicant, a bank tries to predict whether the individual will experience financial distress in the next two years. Hence, from an individual point of view, a positive prediction is less preferable. The prediction is based on 11 biographic and financial history features included in the dataset. There are 18,357 data points. In the simulation, we assume that the world population is finite, consisting of exactly the individuals in the dataset. The distribution is uniform over these individuals, who may change their features strategically, resulting in a new distribution. The original dataset serves as both the initial distribution and the "baseline" distribution d_{BL}, from which modifications are made.

The best response of an individual $(x, y) \in d_{\mathrm{BL}}$ is

$$x_{\mathrm{BR}}(\theta) \xleftarrow{\arg} \max_{x'} u(x', \theta) - c(x', x),$$

where u is the utility function and c is the cost function. The family of classifiers Θ is logistic regression. We use

$$u(x) = -\langle \theta, x \rangle,$$

because a negative value for the utility translates into the more favorable negative prediction. We consider a quadratic cost for feature updates:

$$c(x', x) = \frac{1}{2\varepsilon} \|x' - x\|_2^2.$$

In our experiments, the main parameter we adjust is the sensitivity ε, which controls the strength of strategic response. Additionally, we assume that the individual can change only a subset S of her features, which we call the strategic features. Let x^S be the restriction of x to S. Solving the maximization problem of the individual leads to the to the response

$$x^S_{\mathrm{BR}}(\theta) = x^S - \varepsilon \theta^S.$$

The rest of the features remain unchanged. With that, we can define the strategic response function $\mathcal{D}(\theta)$ as

$$\mathcal{D}(\theta) = \mathrm{Uniform}\left(\{(x_{\mathrm{BR}}(\theta), y) | (x, y) \in d_{\mathrm{BL}}\}\right).$$