Performative Prediction in a Stateful World

Gavin Brown[*]
Boston University

Shlomi Hod[*]
Boston University
[*]equal contribution

Iden Kalemaj[*]
Boston University

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. (ICML 2020). It is an ongoing challenge to understand the influence of such predictions as well as design tools so as to control that influence. We propose a theoretical framework where the response of a target population to the deployed classifier is modeled as a function of the classifier and the current state (distribution) of the population. We show necessary and sufficient conditions for convergence to an equilibrium of two retraining algorithms, *repeated risk minimization* and a lazier variant. Furthermore, convergence is near an optimal classifier. We thus generalize results of Perdomo et al., whose performativity framework does not assume any dependence on the state of the target population. A particular phenomenon captured by our model is that of distinct groups that acquire information and resources at different rates to be able to respond to the latest deployed classifier. We study this phenomenon theoretically and empirically.

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 demonstrate that such predictions can influence the behavior of the target population that they try to predict, e.g., Camacho and Conover (2011); O'neil (2016); Ribeiro et al. (2020). 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 both understand and address the influence of such predictions.

In their recent paper "Performative Prediction," Perdomo, Zrnic, Mendler-Dünner, and Hardt (2020) 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. The equilibrium is a classifier that is optimal for the distribution it induces. Perdomo et al. (2020) model the response of the target population via a deterministic function $\mathcal{D}$ of the published classifier θ. The distribution $\mathcal{D}(\theta)$ 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 bank. These actions depend not only on the bank's classifier, but also on the previous feature state of the individual. As the bank updates its criteria for creditworthiness, new features may become important. Furthermore, when a new classifier is published, different groups in the target population acquire information and adapt their behavior at different rates, so that repeated application of the same classification model may result in continued distribution shifts. We propose a performativity framework that allows for such history dependence.

1.1 Our Framework

We cast the phenomenon of performativity in repeated decision-making as an online learning game (Shalev-Shwartz, 2012). At round t, the institution chooses

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

a classifier θ_t to publish. In response, the adversary 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. Per convention, we also refer to loss as "risk." For now, we set aside 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 responds with

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

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

We call our framework *stateful*, since it incorporates the current distribution of the target population in the performative response map $\texttt{Tr}(;)$, thus possibly preserving information about the state of the population and the history of classifiers played by the institution. Our framework generalizes the *stateless* framework of Perdomo et al. (2020), whose implicit "transition map" depends only on the current classifier θ.

To illustrate our framework, we provide three theoretical examples that are special instances of the model and might be of independent interest. Example 1 and Example 2 capture the particular phenomenon of individuals who act strategically but with outdated information. They provide a starting point for the study of the disparate effects of performativity. We study Example 1 both theoretically and empirically in Section 3 and Section 4, respectively.

Example 1 (k Groups Respond Slowly)**.** Assume there is a deterministic "strategic response function" $\mathcal{D}(\theta)$ unknown to the institution. In response to θ_t, the distribution d_t is a uniform mixture of k distributions $(d_t^{|j})_{j\in[k]}$, one for each group, where

$$d_t^{|j} = \begin{cases} \mathcal{D}(\theta_{t-j+1}) & \text{if } t-j+1 \geq 1 \\ d_0^{|j} & \text{otherwise} \end{cases}. \tag{1}$$

Here $d_0^{|j}$ is the initial distribution for group j. Thus, Group 1 adjusts strategically to the current classifier, while groups with higher indices react to correspondingly older classifiers, modeling a setting where distinct groups receive information at different rates. When $k = 1$ this is the setting of Perdomo et al. (2020).

Example 2 (Geometric Decay Response)**.** As in Example 1, assume that there is a fixed $\mathcal{D}(\theta)$. 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. (2020).

Example 3 (Markov Transitions)**.** To each classifier θ, associate a stochastic matrix A_θ. The transition map is defined as $\texttt{Tr}(d;\theta) = A_\theta d$.

1.2 Our Results

Our goal is to devise a strategy for the institution that converges towards an approximately optimal distribution-classifier pair. Perdomo et al. (2020) 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, akin to retraining heuristics used in practice to deal with different kinds of distribution shifts. Perdomo et al. (2020) analyze RRM in the stateless framework and show that, under convexity and Lipschitz assumptions, it will converge to a near-optimal classifier.

In our setting, when the history of previous classifiers can influence the distribution, it is not clear if such an iterative retraining procedure will succeed. In Example 4, we demonstrate that there are parameter settings for which RRM converges in the stateless model but not in the stateful one.

To control the extent of the performative response, we follow the approach of Perdomo et al. (2020) and impose a Lipschitz requirement on the transition map. It ensures that small changes in the distribution or classifier yield only small changes in the updated distribution. Let Θ denote the set of classifiers, which we assume is a closed convex subset of $\mathbb{R}^d$, and let $\Delta(\mathcal{Z})$ be the space of distributions over examples.

Definition 1 (ε-joint sensitivity)**.** *The transition map* $\texttt{Tr}(;)$ *is* ε*-jointly sensitive if, for all* $\theta, \theta' \in \Theta$ *and* $d, d' \in \Delta(\mathcal{Z})$*,*

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

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

In our model, even repeated deployment of the same classifier can cause "thrashing" behavior, e.g., alternating between two distributions. Therefore, we focus on scenarios where the Lipschitz parameter satisfies $\varepsilon < 1$. In this case, the map $\mathtt{Tr}(\cdot;\theta) : \Delta(\mathcal{Z}) \to \Delta(\mathcal{Z})$ is contractive, and repeated application of the same classifier causes the induced distributions to converge to a *fixed point* that depends only on θ. 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 single response to the institution's classifier.

In the setting where $\varepsilon < 1$, we devise algorithms that converge to an equilibrium pair: a fixed point distribution and a classifier that achieves minimum loss on this distribution. Delayed RRM is a first attempt at an algorithm that converges to an equilibrium pair. It repeats the following: repeatedly deploy the same classifier until we approach a fixed point distribution and only then retrain a classifier that minimizes risk on the current distribution. The goal of Delayed RRM is to overcome the stateful aspect of the population's response by only training on distributions that are close to fixed point distributions. However, it suffers the delay of having to deploy the same classifier for multiple rounds.

Theorem 1 (Informal, see Theorem 8 in Supplement)**.** *If the loss function $\ell(z;\theta)$ is smooth and strongly convex and the transition map $\mathtt{Tr}(d;\theta)$ is Lipschitz in both arguments, then Delayed RRM converges to an equilibrium distribution-classifier pair, coming within distance δ in $O(\log^2 1/\delta)$ rounds.*

As our main result, we show that this delay in retraining is not necessary. We show sufficient conditions, very similar to those in Theorem 1, under which RRM converges to an equilibrium pair. The rate of convergence is much faster than Delayed RRM.

Theorem 2 (Informal, see Theorem 4)**.** *If the loss function $\ell(z;\theta)$ is smooth and strongly convex and the transition map $\mathtt{Tr}(d;\theta)$ is Lipschitz in both arguments, then repeated risk minimization converges to an equilibrium distribution-classifier pair, coming within distance δ in $O(\log 1/\delta)$ rounds.*

As defined, these algorithms require direct access to the data distribution. We also analyze a finite-sample version of RRM and obtain results on the number of datapoints that must be sampled at each round to guarantee linear convergence to the equilibrium pair (see Theorem 5).

For some settings, Delayed RRM may require fewer retraining rounds than RRM until convergence. In Section 3, we compare the two strategies for the scenario of k Groups Respond Slowly (Example 1). This scenario naturally invites a Delayed RRM approach: delay retraining for k rounds until all groups have caught up to the latest classifier. We show that the two algorithms converge to an equilibrium at similar rates in terms of deployment rounds, but if we are only concerned with retraining resources, Delayed RRM is superior.

While converging to an equilibrium pair is a desirable outcome for the institution, this might not be the *optimal* outcome. We formalize the notion of an optimal strategy within our stateful performativity framework. The concept of fixed point distributions is again key to this definition. We show that repeated risk minimization also provides a means to approximate such optimal strategies.

Theorem 3 (Informal, see Theorem 6)**.** *If the loss function $\ell(z;\theta)$ is Lipschitz and strongly convex and the transition map $\mathtt{Tr}(d;\theta)$ is Lipschitz in both arguments, all equilibrium pairs and optimal pairs lie within a small distance of each other.*

Theorem 2 and Theorem 3, which we state formally in Section 2, generalize results of Perdomo et al. (2020) in the stateless framework to our stateful framework. We include complete proofs in the supplementary material.

1.3 Related Work

Our work is closely related to that of Perdomo et al. (2020). Various aspects of the stateless performativity framework of Perdomo et al. (2020) have been studied, such as stochastic and zeroth-order methods for converging to an equilibrium (Drusvyatskiy and Xiao, 2020; Maheshwari et al., 2021; Mendler-Dünner et al., 2020), convergence to the optimal classifier as opposed to an equilibrium (Izzo, Ying, and Zou, 2021; Miller, Perdomo, and Zrnic, 2021), regret minimization (Jagadeesan, Zrnic, and Mendler-Dünner, 2022), and characterization of regions of attraction for different equilibria (Dong and Ratliff, 2021). Narang et al. (2022) propose a multi-player performativity framework where the population reacts to competing institutions' actions. Strategic classification, a term coined by Hardt et al. (2016), is one specific instantiation of performative prediction that has received much attention, e.g., Bechavod et al. (2021); Chen, Liu, and Podimata (2020); Dong and Ratliff (2021); Haghtalab et al. (2020); Hu, Immorlica, and Vaughan (2019); Milli et al. (2019); Munro (2020); Shavit, Edelman, and Axelrod (2020); Tsirtsis and Rodriguez (2020); Zrnic et al. (2021). 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. Hu, Immorlica, and Vaughan (2019) study the disparate effect of strategic manipulation when groups face different costs to manipulation. In k Groups Respond Slowly (Example 1), we model a different aspect of disadvantage, namely access to information.

Prior to our work, Li and Wai (2021) were the only ones to also consider a stateful setting for performative prediction. They study stochastic optimization in the case when the institution samples only one datapoint (or minibatch) at each round, and the updated sample depends both on the current classifier and the previous sample. The samples evolve according to a controlled Markov Chain that depends on the current classifier. The key conceptual difference between the two frameworks is that we update the population-level data based on the previous distribution rather than the realized data from that distribution. Our model subsumes a setting where distributions are updated according to a classifier-dependent Markov chain (see Example 3).

Wood, Bianchin, and Dall'Anese (2021) also propose a more general framework than that of Perdomo et al. (2020), where for every round t a function $\mathcal{D}_t$ maps the classifier to the updated distribution. In contrast to our framework, the updated distributions do not depend on previous distributions. While our framework does not explicitly track time, this can be encoded into the transition function with a simple trick. The focus of Wood, Bianchin, and Dall'Anese (2021) is on stochastic optimization, whereas we mainly obtain population-level results.

Following a preliminary version of our work, Ray et al. (2022) study the geometrically decaying setting of Example 2 and analyze algorithms that converge to the optimal point. In their work, the institution has oracle access (for a fixed batch size) to either the empirical gradient of the loss or the empirical loss function. They provide high probability finite sample guarantees for both types of oracle access. Izzo, Zou, and Ying (2021) study convergence to an optimal point within our framework and provide convergence and sample complexity guarantees, under the assumption that the distributions induced by the classifier are parametric.

Stochastic programming and robust optimization are two general frameworks for modeling optimization problems that involve uncertainty. Within these frameworks, a body of work has studied the case when the system's performance uncertainty depends on the decision variables. Such a setting is usually referred to as decision-dependent distributions or endogenous uncertainty. We refer to Hellemo, Barton, and Tomasgard (2018) and Luo and Mehrotra (2020) and references within for an overview.

Performative prediction can be seen as a special case of reinforcement learning, but the former aims to abstract different phenomena. Performative prediction is designed to capture settings where supervised learning with retraining is a common approach. We explore when natural extensions of supervised techniques to handle performativity (namely RRM) can still succeed and the full strength and complexity of reinforcement learning techniques (such as learning the Q-value function) are unnecessary. For instance, RRM is an algorithm that explicitly only "exploits" and never "explores."

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 $\theta \in \Theta$. In round t, the institution chooses a classifier θ_t. The adversary responds with a distribution $d_t \in \Delta(\mathcal{Z})$ over instances $z = (x, y) \in \mathbb{R}^{m-1} \times \mathbb{R}$ of feature-label pairs. 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.

Algorithm 1 Performative prediction with state

```
Initial distribution $d_0 \in \Delta(\mathcal{Z})$    ▷ Publicly known
for t = 1,2,... do
    Institution publishes $\theta_t \in \Theta$.
    Adversary computes $d_t = \texttt{Tr}(d_{t-1}; \theta_t)$.
    Institution observes $d_t$, suffers loss $\mathbb{E}_{z \sim d_t}[\ell(z; \theta_t)]$.
```

2.1 Repeated Risk Minimization and Stable Pairs

Perdomo et al. (2020) 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 2 (Repeated Risk Minimization (RRM))**.** *Denote by $G(d)$ the risk minimizer*[*]*:*

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

[*] For the scenarios we consider, the set of minimizers will be non-empty. When the set has more than one element, RRM chooses a value from the set arbitrarily.

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

For clarity, we define additional notation wrapping the institution's and adversary's actions into one step.

Definition 3 (RRM map)**.** *Define the RRM map $f\colon \Delta(\mathcal{Z}) \times \Theta \to \Delta(\mathcal{Z}) \times \Theta$ as:*

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

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

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

Definition 5 (Stable Pair)**.** *A distribution-classifier pair $(d_{\mathrm{S}}, \theta_{\mathrm{S}})$ is a* stable pair *if the following hold:*

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

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

Definition 6 (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 7 (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, i.e.,*

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

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

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

Example 4 (RRM may not converge)**.** Take the loss function to be 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 $\mathtt{Tr}(d; \theta) = 1 + \varepsilon d + \varepsilon \theta$, which operates on point mass distributions $d \in [1, \infty)$ of the outcome y. Clearly, the transition function $\mathtt{Tr}$ 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:

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

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

The distance between two successive classifiers 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 θ's produced by RRM fails to converge. When $\varepsilon < \frac{1}{1 + \beta/\gamma} = \frac{1}{2}$, the sequence converges. □

Our main result is sufficient conditions for the convergence of RRM in the stateful framework. Endow the space $\Delta(\mathcal{Z}) \times \Theta$ with the product metric $\mathrm{dist}(\cdot, \cdot)$, so that

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

Theorem 4. *Suppose the transition map $\mathtt{Tr}(;)$ is ε-jointly sensitive and the loss function $\ell(z; \theta)$ is β-jointly smooth and γ-strongly convex. Let $\alpha = \varepsilon(1 + \frac{\beta}{\gamma})$. Then for the RRM map f, and all $d, d' \in \mathcal{Z}$ and $\theta, \theta' \in \Theta$, it holds that*

(a) $\mathrm{dist}(f(d, \theta), f(d', \theta')) \leq \alpha \cdot \mathrm{dist}((d, \theta), (d', \theta'))$.

(b) In particular, if $\alpha < 1$, then f has a unique fixed point which is a stable pair with respect to $\mathtt{Tr}(;)$. RRM will converge to this stable pair at a linear rate.

In the introduction, we discussed Delayed RRM, which is a more complex approach to convergence in the stateful setting. Since RRM converges faster, we keep further discussion of Delayed RRM for the Supplementary Material, where we state and prove Theorem 8.

Finally, we analyze the empirical counterpart of RRM, where at each round the institution only has access to a finite sample from the distribution.

Definition 8 (Repeated Emprical Risk Minimization (RERM))**.** *At timestep t, sample n_t samples from d_{t-1}. Let $\widetilde{d}_{t-1}$ be the uniform distribution on the n_t samples from d_{t-1}. Update the classifier to $\theta_t = G(\widetilde{d}_{t-1})$.*

Theorem 5. *Suppose that the loss $\ell(z, \theta)$ is β-jointly Lipschitz and γ-strongly convex and there exist $\alpha > 1, \mu > 0$ such that $\int_{\mathbb{R}^m} e^{\mu |x|^\alpha} d'(dx)$ is finite for all*

$d' \in \Delta(\mathcal{Z})$. Fix $\delta \in (0,1)$ to be a radius of convergence. Consider running RERM with $n_t = O\Big(\frac{\log(t/p)}{(\varepsilon(1+\frac{\gamma}{\beta})\delta)^m}\Big)$ samples at time t. If the transition map $\texttt{Tr}(;)$ is ε-jointly sensitive and $2\varepsilon\Big(1+\frac{\beta}{\gamma}\Big) < 1$, then with probability $1-p$, the iterates of RERM are within a radius δ of a stable pair for $t \geq \Big(1 - 2\varepsilon\Big(1+\frac{\beta}{\gamma}\Big)\Big)O(\log(1/\delta))$.

Our conditions for convergence of RRM and RERM are similar to, but stricter than, those of the setting of Perdomo et al. (2020). In particular, their results for convergence of RRM only require that $\varepsilon < \frac{\gamma}{\beta}$. Example 4 shows that our requirement is necessary.

2.2 Optimality

Theorem 4 guarantees that RRM converges to an equilibrium, but this stable pair might not be optimal with respect to loss. In fact, it is not obvious how to define the notion of "optimal classifier," since the sequence of distributions played by the adversary depends on the initial distribution. Therefore, to define optimality, we restrict our attention to scenarios where the transition map $\texttt{Tr}(;)$ is ε-jointly sensitive with $\varepsilon < 1$. In this setting, repeatedly applying the same classifier θ causes convergence to a single distribution (Definition 4).

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 9 (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_{OPT} is optimal if it 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_{OPT} for all rounds of classification. We note that if $\texttt{Tr}(;)$ is ε-jointly sensitive with $\varepsilon \geq 1$ then θ_{OPT} may not be defined.

Our definitions of stability and optimality generalize those of Perdomo et al. (2020) for the stateless framework. As pointed out in Perdomo et al. (2020), for a given setting, the optimal classifier does not necessarily coincide with a stable classifier. Our next result shows that RRM approximately approaches optimal classifiers.

Theorem 6. *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 classifier θ_{S} and optimal classifier θ_{OPT} it holds*

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

3 k GROUPS RESPOND SLOWLY

Consider the setting of Example 1, where the target population contains k distinct subpopulations. The j-th subpopulation, for $j \in [k]$, responds strategically to the classifier from j rounds ago. This provides a simple model for investigating performativity in settings where information propagates at different rates. For distribution d, let $d^{|j}$ be the distribution conditioned on being in group j, and denote the mixture by $d = (d^{|1}, d^{|2}, \dots, d^{|k})$. Then the transition function is

$$\texttt{Tr}((d_t^{|1}, \dots, d_t^{|k}), \theta_t) = (\mathcal{D}(\theta_t), d_t^{|1}, \dots, d_t^{|k-1}).$$

We compare two algorithms in this setting: RRM and k-Delayed RRM (Algorithm 2), which in this setting is similar to Delayed RRM. This algorithm updates the classifier θ only every k rounds, after all groups have caught up to the latest deployed classifier and thus have the same distribution.

Algorithm 2 k-Delayed RRM

```
input: number of groups $k \in \mathbb{N}$, initial distribution $d_0 \in \Delta(\mathcal{Z})$
Let $\theta = G(d_0)$; publish $\theta$.
for t = 1,2,... do
    Observe $d_t$.                          ▷ $d_t = \texttt{Tr}(d_{t-1};\theta)$.
    If $t \bmod k = 0$, update $\theta = G(d_t)$.
    Publish $\theta$.
```

Effectively, Algorithm 2 performs RRM on the map $\mathcal{D}(\theta)$ every k rounds, which is the setting of Perdomo et al. (2020). Thus, its rate of convergence is k times of the rate of convergence of RRM in the stateless framework where $k = 1$.

Theorem 7 states the performance of the two algorithms. We note that the transition map defined above is not ε-jointly sensitive for any $\varepsilon < 1$, so our proof of convergence of RRM requires additional analysis beyond Theorem 4. Furthermore, note that the assump-

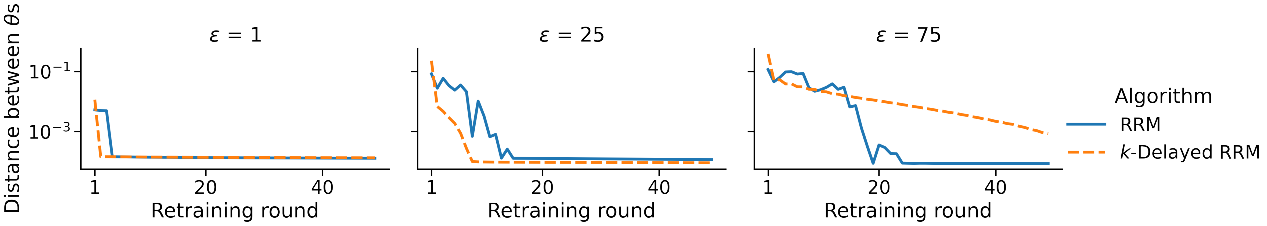


Figure 1: Convergence of RRM and k-Delayed RRM for varying values of ε and $k = 3$. The horizontal axis shows the number of *retraining rounds* and the vertical axis shows the distance between successive θ's.

tion on $\mathcal{D}(\theta)$ corresponds to the notion of *sensitivity* of Perdomo et al. (2020).

Theorem 7. *Suppose that, for some $\varepsilon > 0$ and for all $\theta, \theta' \in \Theta$, the deterministic strategic response map satisfies $\mathcal{W}_1(\mathcal{D}(\theta), \mathcal{D}(\theta')) \leq \varepsilon\|\theta - \theta'\|_2$. Additionally suppose that the loss function $\ell(z;\theta)$ is β-jointly smooth and γ-strongly convex. For $\varepsilon < \frac{\gamma}{\beta}$, the iterates of k-Delayed RRM and RRM converge to θ_{S} at rate $k(1 - \varepsilon\frac{\beta}{\gamma})^t$.*

One advantage of Algorithm 2 is that it can require fewer retraining rounds, possibly requiring fewer computational resources. On the other hand, RRM does not require prior knowledge of the number of groups k.

4 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 (Hardt et al., 2016). It is one instantiation of performative prediction. We adapt a simulation of loan applications in Perdomo et al. (2020), implemented using the `WhyNot` Python package (Miller, Hsu, et al., 2020), and enrich it with state. We demonstrate the convergence of RRM and k-Delayed RRM for the scenario of k Groups Respond Slowly (Example 1) in a credit score setting.[†] We run our experiments on a desktop computer.

The distribution for each of the k groups is deterministically initialized as an instance of the baseline distribution. The baseline distribution is the uniform distribution over samples in Kaggle's *GiveMeSomeCredit* dataset (Kaggle, 2011) consisting of $N = 18,357$ individuals. Hence, there are kN individuals in the whole population.

An individual's strategic response is based on cost and utility functions that take into account the published classifier and the attributes of the individual taken from the *GiveMeSomeCredit* dataset. However, the groups may respond to different θ's at the same round.

The parameter ε controls the strength of the strategic response; larger values allow greater manipulation. We run the simulation for k Groups Respond Slowly using both RRM and k-Delayed RRM, for $k = 3$, $\varepsilon \in \{1, 25, 75\}$, and 300 rounds. Note that for $k = 1$ the two algorithms are identical and the simulation is the same as the stateless setting of Perdomo et al. (2020). Refer to Appendix B.1 for a detailed description of the simulation.

Figure 1 shows the game dynamics when the institution uses the two different algorithms. The vertical axis shows the distance between successively trained classifiers. Recall that in k-Delayed RRM the classifier is retrained once in k rounds (in constrat to RRM where retraining happens every round). Therefore, to depict convergence of the algorithms we plot the distance between successively trained classifiers against the number of *retraining* rounds.

Since the initial distribution consists of k copies of the *GiveMeSomeCredit* dataset, the *same* classifier is deployed at the first retraining round for each ε, and only the most advantaged group responds strategically. For larger ε, the individuals from the most advantaged group respond with a larger update to their features, and therefore in the second round, the institution trains a classifier which is further from the first. In Figure 1 we can see a lower value of $\|\theta_2 - \theta_1\|_2$ for $\varepsilon = 1$ than for the settings of $\varepsilon = 25$ and $\varepsilon = 75$.

For $\varepsilon \in \{1, 25\}$, we see that k-Delayed RRM converges

[†] `https://github.com/shlomihod/performative-prediction-stateful-world`

faster than RRM in terms of number of retraining rounds. Although not depicted here, RRM was superior to k-Delayed RRM in terms of number of rounds (elapsed time) until convergence for all parameter settings we considered. Thus, an important consideration for practitioners in choosing an algorithm is whether elapsed time or retraining resources used until convergence is the more valuable metric.

Interestingly, as demonstrated in Figure 2, for $\varepsilon < 75$ both algorithms reach high accuracy much faster than they converge to the stable pair, and there is little difference between the accuracy dynamics of RRM and k-Delayed RRM. For $\varepsilon = 75$, the accuracy increase of k-Delayed RRM is much slower than that of RRM. In Figure 2, at round $t = 0$ we show the accuracy of the first deployed classifier *before* the strategic response, i.e. the accuracy with respect to the baseline distribution. In all other rounds, the accuracy of the classifier is shown *after* the strategic response.

Finally, we study the game dynamics from the perspective of the target population, stratified by the k groups. Due to the different rate of information acquisition, there is a hierarchy of advantage among the groups. Intuitively, the group that responds first to the latest deployed classifier has an *advantage*[‡] compared to other groups. We investigate if this hierarchy of advantage translates into a disparity between groups in the benefit they achieve from their strategic response. This raises a question of whether the choice of algorithm can mitigate the disparity effect. We conjecture that since k-Delayed RRM allows the groups to reach a homogeneous distribution, it is preferable to RRM in terms of per-group disadvantage.

We investigate the per-group *negative rate* (NR), the proportion of individuals that are predicted unlikely to default, as a measure of the groups' benefit from their strategic response. We focus on $\varepsilon = 25$ in our analysis. Figure 3 shows that in the initial rounds ($1 \leq t < 10$), before NR converges to the same value for all groups, there is a noticeable difference between the groups for both algorithms. When accumulated over the first 10 rounds, the most advantaged group has 2.37% and 2.09% more negative predictions than the most disadvantaged group for RRM and k-Delayed RRM, respectively. It appears that in this simulation there is no clear advantage to either one of the algorithms in terms of per-group disparity.

The definition of k Groups Respond Slowly allows for different initial distribution for each group. In Appendix B.2, we run additional experiments with varying initial distributions for the groups to supplement our existing results.

5 DISCUSSION

This work models the important role that the history of institutional predictions plays in shaping the behavior of individuals. The addition of state to the formal model of performativity opens up a new venue for discussing the social impact of machine 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. We believe the setting of Example 1, k Groups Respond Slowly, provides an excellent starting point for future investigation of the interaction between performativity and fairness in the presence of information disparities.

The theoretical assumptions in this work constrain either the loss function or the transition function. While an institution controls the loss function, there is no clear way to verify in practice the existence of a fixed, deterministic transition function, let alone its sensitivity. Our theoretical guarantees are thus best interpreted in practice as conditional statements: if the world's response is not too sensitive to prior history or the institution's choices, the simple and widely used heuristic of RRM is sensible.

This work focuses on the goal of convergence, but if convergence is not a priority for the institution then it may be interesting to study which measures of the institution'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.

6 SOCIETAL 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.

Acknowledgements

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

[‡] We stress that the operationalization of disadvantage and access to information is deliberately abstracted in our model, and we do not attempt to fully capture these complex societal concepts.

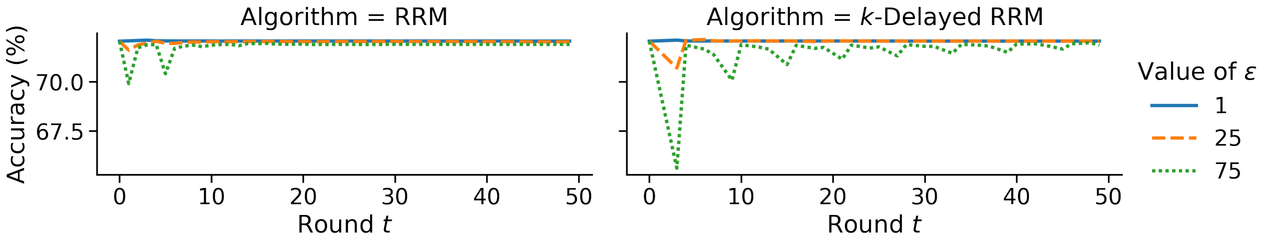


Figure 2: Accuracy of RRM and k-Delayed RRM for different ε and $k = 3$. The horizontal axis shows the number of *rounds* and the vertical axis shows the accuracy of the published model after the strategic response.

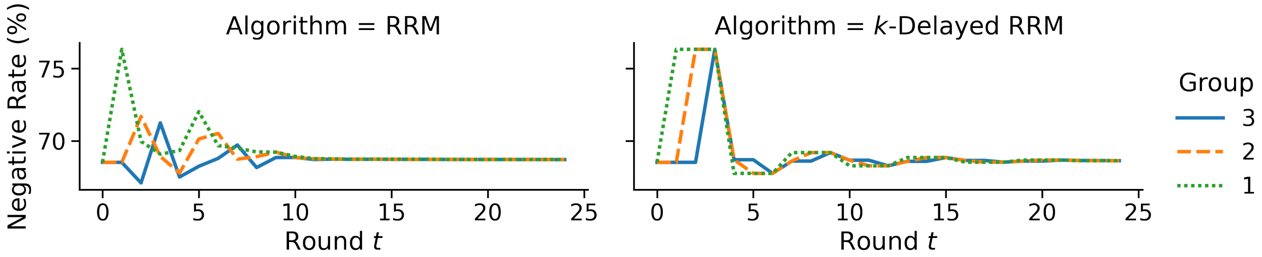


Figure 3: Negative rate of RRM and k-Delayed RRM for $\varepsilon = 25$ and $k = 3$. The horizontal axis shows the number of *rounds* and the vertical axis shows the negative rate of the published model after the strategic response.

References

Bechavod, Yahav et al. (2021). “Gaming Helps! Learning from Strategic Interactions in Natural Dynamics”. In: *Proceedings, International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1234–1242.

Bubeck, Sébastien (2015). “Convex Optimization: Algorithms and Complexity”. In: *Foundations and Trends in Machine Learning* 8(3-4), pp. 231–357.

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

Chen, Yiling, Yang Liu, and Chara Podimata (2020). “Learning Strategy-Aware Linear Classifiers”. In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Dong, Roy and Lillian J. Ratliff (2021). “Which Echo Chamber? Regions of Attraction in Learning with Decision-Dependent Distributions”. In: *CoRR* abs/2107.00055.

Drusvyatskiy, Dmitriy and Lin Xiao (2020). “Stochastic optimization with decision-dependent distributions”. In: *CoRR* abs/2011.11173.

Fournier, Nicolas and Arnaud Guillin (2015). “On the Rate of Convergence in Wasserstein Distance of the Empirical Measure”. In: *Probability Theory and Related Fields* 162 (3), pp. 709–738.

Haghtalab, Nika et al. (2020). “Maximizing Welfare with Incentive-Aware Evaluation Mechanisms”. In: *Proceedings, International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 160–166.

Hardt, Moritz et al. (2016). “Strategic classification”. In: *Proceedings of the ACM conference on Innovations in Theoretical Computer Science (ITCS)*, pp. 111–122.

Hellemo, Lars, Paul I. Barton, and Asgeir Tomasgard (2018). “Decision-dependent probabilities in stochastic programs with recourse”. In: *Computational Management Science* 15(3-4), pp. 369–395.

Hu, Lily, Nicole Immorlica, and Jennifer Wortman Vaughan (2019). “The Disparate Effects of Strategic Manipulation”. In: *Proceedings, Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 259–268.

Izzo, Zachary, Lexing Ying, and James Zou (2021). “How to Learn when Data Reacts to Your Model: Performative Gradient Descent”. In: *Proceedings, International Conference on Machine Learning (ICML)*, pp. 4641–4650.

Izzo, Zachary, James Zou, and Lexing Ying (2021). "How to Learn when Data Gradually Reacts to Your Model". In: *CoRR* abs/2112.07042.

Jagadeesan, Meena, Tijana Zrnic, and Celestine Mendler-Dünner (2022). "Regret Minimization with Performative Feedback". In: *CoRR* abs/2202.00628.

Kaggle (2011). *Give Me Some Credit Dataset*. `https://www.kaggle.com/c/GiveMeSomeCredit`.

Li, Qiang and Hoi-To Wai (2021). "State Dependent Performative Prediction with Stochastic Approximation". In: *CoRR* abs/2110.00800.

Luo, Fengqiao and Sanjay Mehrotra (2020). "Distributionally robust optimization with decision dependent ambiguity sets". In: *Optimization Letters* 14(8), pp. 2565–2594.

Maheshwari, Chinmay et al. (2021). "Zeroth-Order Methods for Convex-Concave Minmax Problems: Applications to Decision-Dependent Risk Minimization". In: *CoRR* abs/2106.09082.

Mendler-Dünner, Celestine et al. (2020). "Stochastic Optimization for Performative Prediction". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Miller, John, Chloe Hsu, et al. (2020). *WhyNot*.

Miller, John, Juan C. Perdomo, and Tijana Zrnic (2021). "Outside the Echo Chamber: Optimizing the Performative Risk". In: *Proceedings, International Conference on Machine Learning (ICML)*, pp. 7710–7720.

Milli, Smitha et al. (2019). "The Social Cost of Strategic Classification". In: *Proceedings, Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 230–239.

Munro, Evan (2020). "Learning to Personalize Treatments When Agents Are Strategic". In: *CoRR* abs/2011.06528.

Narang, Adhyyan et al. (2022). "Multiplayer Performative Prediction: Learning in Decision-Dependent Games". In: *CoRR* abs/2201.03398.

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

Perdomo, Juan C. et al. (2020). "Performative Prediction". In: *Proceedings, International Conference on Machine Learning (ICML)*, pp. 7599–7609.

Ray, Mitas et al. (2022). "Decision-Dependent Risk Minimization in Geometrically Decaying Dynamic Environments". In: *AAAI Conference on Artificial Intelligence*, To appear.

Ribeiro, Manoel Horta et al. (2020). "Auditing radicalization pathways on Youtube". In: *Proceedings, Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 131–141.

Shalev-Shwartz, Shai (2012). "Online Learning and Online Convex Optimization". In: *Foundations and Trends in Machine Learning* 4(2), pp. 107–194.

Shavit, Yonadav, Benjamin L. Edelman, and Brian Axelrod (2020). "Causal Strategic Linear Regression". In: *Proceedings, International Conference on Machine Learning (ICML)*, pp. 8676–8686.

Tsirtsis, Stratis and Manuel Gomez Rodriguez (2020). "Decisions, Counterfactual Explanations and Strategic Behavior". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Wood, Killian, Gianluca Bianchin, and Emiliano Dall'Anese (2021). "Online Projected Gradient Descent for Stochastic Optimization with Decision-Dependent Distributions". In: *CoRR* abs/2107.09721.

Zrnic, Tijana et al. (2021). "Who Leads and Who Follows in Strategic Classification?" In: *CoRR* abs/2106.12529.

Supplementary Material: Performative Prediction in a Stateful World

A PROOFS OF THEOREMS

We first state some key lemmas used in the proofs of our theorems. The second lemma follows directly from the proof of Theorem 3.5 of Perdomo et al. (2020).

Lemma 1 (Bubeck (2015))**.** *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 $(y - x_*)^\top \nabla g(x_*) \geq 0$ for all $y \in \Omega$.

Lemma 2 (Perdomo et al. (2020))**.** *Suppose the loss function $\ell(z;\theta)$ is β-jointly smooth and γ-strongly-convex. For any two distributions $d, d' \in \Delta(\mathcal{Z})$, risk minimization satisfies*

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

Lemma 3. *Suppose the map $\texttt{Tr}(;)$ is ε-jointly sensitive with $\varepsilon < 1$. Then for any $\theta_1, \theta_2 \in \Theta$ 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. By the definition of fixed point distributions, $\mathcal{W}_1(d_1, d_2) = \mathcal{W}_1(\texttt{Tr}(d_1, \theta_1), \texttt{Tr}(d_2, \theta_2))$. Since the transition map is ε-jointly sensitive, we obtain

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

The lemma follows from the equation above. □

A.1 Proof of Theorem 1

In this section, we formally state our Delayed RRM algorithm (Algorithm 3) and prove Theorem 1, restated formally below in Theorem 8.

Algorithm 3 Main procedure **Delayed RRM**

```
Require: radius $\delta$, loss function $\ell$, initial distribution $d_0$, sensitivity $\varepsilon \in (0,1)$
Let $d \leftarrow d_0$.
Initialize $\theta \leftarrow \mathbf{0}$.
loop $t$ times:
    Calculate $\theta = \arg\min_{\theta'} \mathbb{E}_{z \sim d} \ell(z;\theta')$.
    Update $d \leftarrow$ RepeatedDeployment$(d, \theta, \delta, \varepsilon)$.
```

Algorithm 4 Sub-procedure **RepeatedDeployment**

```
Require: initial distribution $d_0$, classifier $\theta$, radius $\delta$, sensitivity $\varepsilon \in (0,1)$.
Publish $\theta$.
Observe $d_1 \leftarrow \mathtt{Tr}(d_0;\theta)$.
Let $d_{\text{current}} \leftarrow d_1$.
Let $r = \log^{-1}(\frac{1}{\varepsilon})\log(\frac{\mathcal{W}_1(d_0,d_1)}{\delta})$.
loop $r$ times:
    Publish $\theta$.
    Observe $d_{\text{current}} \leftarrow \mathtt{Tr}(d_{\text{current}};\theta)$.
Return $d_{\text{current}}$.
```

Theorem 8. *Suppose the loss function $\ell(z;\theta)$ is β-jointly smooth and γ-strongly convex. Let $\theta_1,\dots,\theta_t$ denote the classifiers obtained in each iteration of Step 4 of Algorithm 3. If the transition map $\mathtt{Tr}(\cdot;\cdot)$ is ε-jointly sensitive with $\varepsilon < 1$ and $\frac{\varepsilon}{1-\varepsilon} < \frac{\gamma}{2\beta}$, then:*

(a) $\|\theta_t - \theta_{\mathrm{S}}\|_2 \le \delta$ for $t \ge \big(1 - \frac{2\varepsilon\beta}{\gamma(1-\varepsilon)}\big)^{-1}\log(\frac{\|\theta_0-\theta_{\mathrm{S}}\|_2}{\delta})$.

(b) Each iteration of the main **loop** *in Algorithm 3 consists of $r \le \log(\frac{1}{\varepsilon})^{-1}\log(\frac{\operatorname{diam}(\Delta(\mathcal{Z}))}{\delta})$ deployments of the same classifier, where $\operatorname{diam}(\Delta(\mathcal{Z}))$ denotes the largest $\mathcal{W}_1$ distance between two distributions in $\Delta(\mathcal{Z})$. Thus, $O(\log^2(\frac{1}{\delta}))$ deployments are needed for θ_t to be within distance δ of θ_{S}.*

We first prove an auxiliary lemma regarding the **RepeatedDeployment** procedure.

Lemma 4. *Given a classifier θ, denote by $\widetilde{d}_\theta$ the distribution returned from* **RepeatedDeployement**$(d_0,\theta,\delta,\varepsilon)$ *for any $d_0 \in \Delta(\mathcal{Z})$. Let d_θ be the fixed point distribution for θ. If $\varepsilon < 1$, then*

$$\mathcal{W}_1(\widetilde{d}_\theta, d_\theta) \le \frac{\varepsilon}{1-\varepsilon}\delta.$$

Proof. For a fixed θ and $\varepsilon < 1$, the map $\mathtt{Tr}(\cdot;\theta)$ is contractive with Lipschitz coefficient ε and has a unique fixed point d_θ. Note that $\widetilde{d}_\theta = \mathtt{Tr}^{r+1}(d_0;\theta)$, where the transition map is applied $r+1$ times with the same classifier θ. Let $d_1 = \mathtt{Tr}(d_0;\theta)$. It is easy to see that

$$\mathcal{W}_1(\widetilde{d}_\theta, d_\theta) \le \frac{\varepsilon^{r+1}}{1-\varepsilon}\mathcal{W}_1(d_0,d_1). \tag{2}$$

For $r = \log^{-1}(\frac{1}{\varepsilon})\log(\frac{\mathcal{W}_1(d_0,d_1)}{\delta})$, we obtain $\varepsilon^r = \frac{\delta}{\mathcal{W}_1(d_0,d_1)}$. Plugging the value of ε^r into (2) concludes the proof. □

Proof of Theorem 8. We prove part (a). As before, let $\widetilde{d}_\theta$ denote the distribution returned from a single call of **RepeatedDeployement**$(\cdot,\theta,\delta,\varepsilon)$. Note that

$$\theta_{i+1} = G(\widetilde{d}_{\theta_i}).$$

Recall that $\theta_{\mathrm{S}} = G(d_{\mathrm{S}})$. Then, Lemma 2 gives

$$\|\theta_{i+1} - \theta_{\mathrm{S}}\|_2 = \|G(\widetilde{d}_{\theta_i}) - G(d_{PS})\|_2 \le \frac{\beta}{\gamma}\cdot\mathcal{W}_1(\widetilde{d}_{\theta_i}, d_{\mathrm{S}}). \tag{3}$$

We bound the distance $\mathcal{W}_1(\widetilde{d}_{\theta_i}, d_{\mathrm{S}})$. Let d_{θ_i} be the fixed point distribution for θ_i. By the triangle inequality,

$$\mathcal{W}_1(\widetilde{d}_{\theta_i}, d_{\mathrm{S}}) \le \mathcal{W}_1(\widetilde{d}_{\theta_i}, d_{\theta_i}) + \mathcal{W}_1(d_{\theta_i}, d_{\mathrm{S}}). \tag{4}$$

The first term of the sum in (4) can be bound using Lemma 4, whereas the second term can be bounded using Lemma 3 and the fact that d_{θ_i} and d_{S} are fixed point distributions. We obtain

$$\mathcal{W}_1(\widetilde{d}_{\theta_i}, d_{\mathrm{S}}) \le \frac{\varepsilon}{1-\varepsilon}\delta + \frac{\varepsilon}{1-\varepsilon}\|\theta_i - \theta_{\mathrm{S}}\|_2. \tag{5}$$

We consider two cases. In the case when $\|\theta_i - \theta_{\mathrm{S}}\|_2 > \delta$, we show that in the next retraining round, the classifier θ_{i+1} moves closer to θ_{S}. Replace $\delta < \|\theta_i - \theta_{\mathrm{S}}\|_2$ in (5). Then from (3) we obtain $\|\theta_{i+1} - \theta_{\mathrm{S}}\|_2 \leq \frac{2\varepsilon}{1-\varepsilon}\frac{\beta}{\gamma}\|\theta_i - \theta_{\mathrm{S}}\|_2$, i.e., contraction happens in iteration $i+1$. In the case when $\|\theta_i - \theta_{\mathrm{S}}\|_2 \leq \delta$, combining (3) and (5) yields

$$\|\theta_{i+1} - \theta_{\mathrm{S}}\|_2 \leq \frac{2\varepsilon}{1-\varepsilon}\frac{\beta}{\gamma}\delta \leq \delta.$$

This shows that the classifier θ_{i+1} does not leave the ball of radius δ around θ_{S}. The two cases combined give that for $t \geq \left(1 - \frac{2\varepsilon\beta}{\gamma(1-\varepsilon)}\right)^{-1}\log(\frac{\|\theta_0-\theta_{\mathrm{S}}\|_2}{\delta})$ iterations we have

$$\|\theta_t - \theta_{PS}\|_2 \leq \left(\frac{2\varepsilon\beta}{\gamma(1-\varepsilon)}\right)^t \|\theta_0 - \theta_{\mathrm{S}}\|_2 \leq \delta,$$

which concludes the proof of part (a). Part (b) is clear from the statement of the subprocedure **RepeatedDeployement**. □

A.2 Proof of Theorem 4

Proof of Theorem 4. Note that if part (a) holds, then part (b) follows from the fact that the map f is contractive with contraction coefficient $\varepsilon(1+\frac{\beta}{\gamma}) < 1$. 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 pair for it satisfies $d^* = \mathtt{Tr}(d^*, \theta^*)$ and $\theta^* = G(\mathtt{Tr}(d^*, \theta^*)) = G(d^*)$.

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

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

The ε-joint sensitivity of the transition map yields

$$\mathcal{W}_1(\mathtt{Tr}(d,\theta), \mathtt{Tr}(d',\theta')) \leq \varepsilon\mathcal{W}_1(d,d') + \varepsilon\|\theta - \theta'\|_2. \tag{6}$$

We will show that

$$\|G(d,\theta) - G(d',\theta')\|_2 \leq \varepsilon\frac{\beta}{\gamma}\mathcal{W}_1(d,d') + \varepsilon\frac{\beta}{\gamma}\|\theta - \theta'\|_2. \tag{7}$$

Combining equations (6) and (7) will conclude the proof. We obtain (7) by using Lemma 2 with distributions $\mathtt{Tr}(d,\theta)$ and $\mathtt{Tr}(d',\theta')$ together with the sensitivity of the transition map. This gives

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

□

A.3 Proof of Theorem 5

In this section, we show Theorem 5 on the performance of RERM. In particular, we show that for iterates (d_{t-1}, θ_t) of RERM it holds that

$$\mathrm{dist}((d_{t-1},\theta_t),(d_{\mathrm{S}},\theta_{\mathrm{S}})) \leq \delta \text{ for all } t \geq \Big(1 - 2\varepsilon\Big(1+\frac{\beta}{\gamma}\Big)\Big)\log\Big(\frac{\mathrm{dist}((d_0,\theta_1),(d_{\mathrm{S}},\theta_{\mathrm{S}}))}{\delta}\Big).$$

Proof of Theorem 5. Given a distribution d, let $d^{(n)}$ denote the empirical distribution over the n samples drawn from d. Let $\widetilde{G}(d) := G(d^{(n)})$. Define the RERM Map $\widetilde{f}(d,\theta) = (\mathtt{Tr}(d,\theta), \widetilde{G}(\mathtt{Tr}(d,\theta))$. Our analysis relies on the following assumption.

Assumption 1. *For each timestep $t \geq 1$, it holds that $\mathcal{W}_1(d_t^{(n_t)}, d_t) \leq \varepsilon\Big(1+\frac{\gamma}{\beta}\Big)\delta$.*

Given Assumption 1, we show that one of the following holds for each $t \geq 1$:

1. If the iterate (d_{t-1}, θ_t) is at distance at least δ from $(d_\mathrm{S}, \theta_\mathrm{S})$, then the distance of the next iterate (d_t, θ_{t+1}) to $(d_\mathrm{S}, \theta_\mathrm{S})$ will contract by a factor of at least $2\varepsilon\Big(1+\frac{\beta}{\gamma}\Big)$.

2. If the iterate (d_{t-1}, θ_t) is within a ball of radius δ of $(d_\mathrm{S}, \theta_\mathrm{S})$, then the next iterate will also be within this ball.

By Theorem 2 of Fournier and Guillin (2015), if $n_t = O\Big(\frac{1}{(\varepsilon(1+\frac{\gamma}{\beta})\delta)^m}\log(t/p)\Big)$, then $\mathcal{W}_1(d_t^{(n_t)}, d_t) \geq \varepsilon\Big(1+\frac{\gamma}{\beta}\Big)\delta$ with probability at most $\frac{6p}{\pi^2 t^2}$. By a Union Bound over all t, we obtain that Assumption 1 holds with probability at least $1 - \sum_{t=1}^{\infty}\frac{6p}{\pi^2 t^2} = 1 - p$.

We start by showing Item 1 is true under Assumption 1. Suppose $\operatorname{dist}((d,\theta),(d_\mathrm{S},\theta_\mathrm{S})) \geq \delta$. Then

$$\operatorname{dist}(\widetilde{f}(d,\theta),(d_\mathrm{S},\theta_\mathrm{S}))$$

$$= \mathcal{W}_1(\mathtt{Tr}(d,\theta), d_\mathrm{S}) + \|\widetilde{G}(\mathtt{Tr}(d,\theta)) - \theta_\mathrm{S}\|_2$$

$$\leq \mathcal{W}_1(\mathtt{Tr}(d,\theta), \mathtt{Tr}(d_\mathrm{S},\theta_\mathrm{S})) + \|\widetilde{G}(\mathtt{Tr}(d,\theta)) - G(\mathtt{Tr}(d,\theta))\|_2 + \|G(\mathtt{Tr}(d,\theta)) - G(\mathtt{Tr}(d_\mathrm{S},\theta_\mathrm{S}))\|_2 \tag{8}$$

$$\leq \varepsilon\mathcal{W}_1(d, d_\mathrm{S}) + \varepsilon\left\|\theta-\theta_\mathrm{S}\right\|_2 + \frac{\beta}{\gamma}\mathcal{W}_1(\mathtt{Tr}(d,\theta)^{(n)}, \mathtt{Tr}(d,\theta)) + \frac{\beta}{\gamma}\mathcal{W}_1(\mathtt{Tr}(d,\theta), \mathtt{Tr}(d_\mathrm{S},\theta_\mathrm{S})) \tag{9}$$

$$\leq \varepsilon\mathcal{W}_1(d, d_\mathrm{S}) + \varepsilon\left\|\theta-\theta_\mathrm{S}\right\|_2 + \frac{\beta}{\gamma}\varepsilon\Big(1+\frac{\gamma}{\beta}\Big)\delta + \frac{\beta}{\gamma}(\varepsilon\mathcal{W}_1(d, d_\mathrm{S}) + \varepsilon\left\|\theta-\theta_\mathrm{S}\right\|_2) \tag{10}$$

$$= \varepsilon\Big(1+\frac{\beta}{\gamma}\Big)\operatorname{dist}((d,\theta),(d_\mathrm{S},\theta_\mathrm{S})) + \varepsilon\Big(1+\frac{\beta}{\gamma}\Big)\delta$$

$$\leq 2\varepsilon\Big(1+\frac{\beta}{\gamma}\Big)\operatorname{dist}((d,\theta),(d_\mathrm{S},\theta_\mathrm{S})). \tag{11}$$

In Eq. (8) we use the triangle inequality. Eq. (9) follows by applying the ε-Lipschitzness of the transition map to bound the first term and Lemma 2 to bound the second and third term. In Eq. (10), we use Assumption 1 to bound the second term and ε-Lipschitzness to bound the third term. Finally, in Eq. (11), we use the assumption that $\operatorname{dist}((d,\theta),(d_\mathrm{S},\theta_\mathrm{S})) \geq \delta$. Therefore, Item 1 holds.

To show Item 2, suppose that $\operatorname{dist}((d,\theta),(d_\mathrm{S},\theta_\mathrm{S})) < \delta$. We show that $\operatorname{dist}(\widetilde{f}(d,\theta),(d_\mathrm{S},\theta_\mathrm{S})) < \delta$. From the previous argument, under Assumption 1, it holds

$$\operatorname{dist}(\widetilde{f}(d,\theta),(d_\mathrm{S},\theta_\mathrm{S})) \leq \varepsilon\Big(1+\frac{\beta}{\gamma}\Big)\operatorname{dist}((d,\theta),(d_\mathrm{S},\theta_\mathrm{S})) + \varepsilon\Big(1+\frac{\beta}{\gamma}\Big)\delta \leq 2\varepsilon\Big(1+\frac{\beta}{\gamma}\Big)\delta.$$

From the assumption that $2\varepsilon\Big(1+\frac{\beta}{\gamma}\Big) < 1$ we obtain Item 2.

It remains to show that under Assumption 1 the iterates (d_{t-1}, θ_t) will reach a ball of radius δ around $(d_\mathrm{S}, \theta_\mathrm{S})$ for $t \geq \Big(1 - 2\varepsilon\Big(1+\frac{\beta}{\gamma}\Big)\Big)\log\Big(\frac{\operatorname{dist}((d_0,\theta_1),(d_\mathrm{S},\theta_\mathrm{S}))}{\delta}\Big)$. Suppose the assumption on t holds. Furthermore, suppose that none of the iterates $(d_0, \theta_1), \ldots, (d_{t-2}, \theta_{t-1})$ are within a radius δ of $(d_\mathrm{S}, \theta_\mathrm{S})$. By Assumption 1, Item 1, it holds

$$\begin{aligned}\operatorname{dist}((d_{t-1},\theta_t),(d_\mathrm{S},\theta_\mathrm{S})) &\leq \Big(2\varepsilon\Big(1+\frac{\beta}{\gamma}\Big)\Big)^t\operatorname{dist}((d_0,\theta_1),(d_\mathrm{S},\theta_\mathrm{S}))\\ &\leq \exp\Big(-t\Big(1-2\varepsilon\Big(1+\frac{\beta}{\gamma}\Big)\Big)\Big)\operatorname{dist}((d_0,\theta_1),(d_\mathrm{S},\theta_\mathrm{S}))\\ &\leq \delta.\end{aligned}$$

If one of the iterates $(d_0, \theta_1), \ldots, (d_{t-2}, \theta_{t-1})$ is within a radius δ of $(d_\mathrm{S}, \theta_\mathrm{S})$, then by Assumption 1, Item 2, all consecutive iterates will also be within a radius δ of a stable pair. Since Assumption 1 holds with probability $1-p$, this concludes the proof. □

A.4 Proof of Theorem 6

Proof of Theorem 6. Our proof is similar to an argument of Perdomo et al. (2020). Let θ_OPT be an optimal classifier and let d_OPT be its corresponding fixed point distribution. Let θ_S be a stable classifier, with corresponding fixed point distribution d_{S}. By the definitions of optimality and stability,

$$\mathop{\mathbb{E}}_{z\sim d_{\mathrm{OPT}}} \ell(z;\theta_{\mathrm{OPT}}) \leq \mathop{\mathbb{E}}_{z\sim d_{\mathrm{S}}} \ell(z;\theta_{\mathrm{S}}) \leq \mathop{\mathbb{E}}_{z\sim d_{\mathrm{S}}} \ell(z;\theta_{\mathrm{OPT}})\,. \tag{12}$$

We first show that

$$\mathop{\mathbb{E}}_{z\sim d_{\mathrm{S}}} \ell(z;\theta_{\mathrm{OPT}}) - \mathop{\mathbb{E}}_{z\sim d_{\mathrm{S}}} \ell(z;\theta_{\mathrm{S}}) \geq \frac{\gamma}{2}\|\theta_{\mathrm{OPT}}-\theta_{\mathrm{S}}\|_2^2\,. \tag{13}$$

By the strong convexity of the loss function, for all $z \in \mathcal{Z}$ it holds

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

As a result,

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

Since θ_{S} minimizes the value of ℓ over the distribution d_{S}, Lemma 1 implies

$$\mathop{\mathbb{E}}_{z\sim d_{\mathrm{S}}}\big[\nabla_\theta \ell(z;\theta_{\mathrm{S}})^\top(\theta_{\mathrm{OPT}}-\theta_{\mathrm{S}})\big] \geq 0.$$

Therefore, Eq. (13) holds. On the other hand, since the loss is L_z-Lipschitz in z, by Lemma 3,

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

If $\frac{\varepsilon}{1-\varepsilon} < \frac{\gamma\|\theta_{\mathrm{OPT}}-\theta_{\mathrm{S}}\|_2}{2L_z}$ then $\frac{L_z\varepsilon}{1-\varepsilon}\|\theta_{\mathrm{OPT}}-\theta_{\mathrm{S}}\|_2 \leq \frac{\gamma}{2}\|\theta_{\mathrm{OPT}}-\theta_{\mathrm{S}}\|_2^2$. This would imply that

$$\mathop{\mathbb{E}}_{z\sim d_{\mathrm{S}}} \ell(z;\theta_{\mathrm{OPT}}) - \mathop{\mathbb{E}}_{z\sim d_{\mathrm{OPT}}} \ell(z;\theta_{\mathrm{OPT}}) \leq \mathop{\mathbb{E}}_{z\sim d_{\mathrm{S}}} \ell(z;\theta_{\mathrm{OPT}}) - \mathop{\mathbb{E}}_{z\sim d_{\mathrm{S}}} \ell(z;\theta_{\mathrm{S}}),$$

which contradicts Eq. (12). Therefore, $\frac{\varepsilon}{1-\varepsilon} \geq \frac{\gamma\|\theta_{\mathrm{OPT}}-\theta_{\mathrm{S}}\|_2}{2L_z}$, as desired. □

A.5 Proof of Theorem 7

In this section, we show Theorem 7 on the performance of RRM and k-Delayed RRM for the setting of Example 1.

Proof of Theorem 7. **k-Delayed RRM.** Let (d_{kt-1}, θ_{kt}) be the iterates of Algorithm 2 at timestep kt for $t \geq 1$. Since kt is a multiple of k, then $\theta_{kt} = G(d_{kt-1})$ and all components $d_{kt}^{|j}$, $j \in [k]$ of the mixture distribution d_{kt-1} are identical. Therefore $\theta_{kt} = G(d_{kt-1}^{|k})$. Additionally, $d_{k(t+1)-1}^{|k} = \mathcal{D}(\theta_{kt})$. Therefore, the iterates $(d_{kt-1}^{|k}, \theta_{kt})$, where $t \geq 1$, correspond to the iterates of RRM in the stateless setting of Perdomo et al. (2020). By Theorem 3.5 of Perdomo et al. (2020), the iterates θ_{kt} converge to θ_{S} at rate $\left(1-\varepsilon\frac{\beta}{\gamma}\right)^t$. In turn, it follows that the iterates of Algorithm 2 converge to θ_{S} at rate $k\left(1-\varepsilon\frac{\beta}{\gamma}\right)^t$.

RRM. Let θ_{S} denote a stable classifier for $\mathcal{D}(\theta)$, and let $d_{\mathrm{S}} = \mathcal{D}(\theta_{\mathrm{S}})$. Let $\mathbf{d}_{\mathrm{S}} = (d_{\mathrm{S}}, \dots, d_{\mathrm{S}})$ be the uniform mixture on k identical components d_{S}. Then, $(\mathbf{d}_{\mathrm{S}}, \theta_{\mathrm{S}})$ is a stable point for $\texttt{Tr}$ because $\texttt{Tr}(\mathbf{d}_{\mathrm{S}}, \theta_{\mathrm{S}}) = (\mathcal{D}(\theta_{\mathrm{S}}), d_{\mathrm{S}}, \dots, d_{\mathrm{S}}) = \mathbf{d}_{\mathrm{S}}$, and θ_{S} is the classifier that minimizes loss on $\mathbf{d}_{\mathrm{S}}$. Let (d_{t-1}, θ_t) denote the iterates of RRM in the setting of Example 1 and suppose $t \geq k$. Then,

$$\begin{aligned}
\|\theta_t - \theta_{\mathrm{S}}\|_2 = \|G(d_{t-1}) - G(\theta_{\mathrm{S}})\|_2 &\leq \frac{\beta}{\gamma}\mathcal{W}_1(d_{t-1}, \mathbf{d}_{\mathrm{S}}) \\
&\leq \frac{\beta}{\gamma}\frac{1}{k}\sum_{i=1}^{k}\mathcal{W}_1(d_{t-1}^{|i}, d_{\mathrm{S}}) \\
&= \frac{\beta}{\gamma}\frac{1}{k}\sum_{i=1}^{k}\mathcal{W}_1(\mathcal{D}(\theta_{t-i}), \mathcal{D}(\theta_{\mathrm{S}})) \\
&\leq \frac{\beta}{\gamma}\frac{1}{k}\sum_{i=1}^{k}\|\theta_{t-i} - \theta_S\|_2\,,
\end{aligned}$$

where the first inequality follows by Lemma 2, the second inequality follows by the definition of d_{t-1} as a uniform mixture on k components, and the third inequality follows by the ε-sensitivity of the map $\mathcal{D}(\cdot)$.

Let $\delta_t = \|\theta_t - \theta_{\mathrm{S}}\|_2$. To find the rate of convergence of the sequence $\delta_t, t \geq 1$, it suffices to find the rate of convergence for the sequence that satisfies the following linear recurrence

$$\delta_t = \frac{\varepsilon\beta}{\gamma} \cdot \frac{1}{k}(\delta_{t-1} + \cdots + \delta_{t-k}).$$

The sequence decays exponentially with a rate the depends on $\varepsilon, \beta, \gamma, k$. We provide a simpler analysis, using the inequality

$$\delta_t \leq \frac{\varepsilon\beta}{\gamma} \max_{i \in [k]} \delta_{t-i}.$$

For $j \geq 0$, let $i_j^* = \arg\max_{i\in[k]} \delta_{jk+i}$. Define $\tilde{\theta}_j = \theta_{i_j^*}$. By the above argument, $\delta_{jk+1} \leq \frac{\varepsilon\beta}{\gamma} \max_{i\in[k]} \delta_{jk+1-i} = \max_{\ell\in[k]} \delta_{(j-1)k+\ell} = \frac{\varepsilon\beta}{\gamma} \left\|\tilde{\theta}_{j-1} - \theta_{\mathrm{S}}\right\|_2$.

Now, $\delta_{jk+2} \leq \frac{\varepsilon\beta}{\gamma} \max_{i\in[k]} \delta_{jk+2-i} \leq \frac{\varepsilon\beta}{\gamma} \max_{i\in[k+1]} \delta_{jk+2-i} \leq \frac{\varepsilon\beta}{\gamma} \max\{\delta_{jk+1}, \max_{\ell\in[k]} \delta_{(j-1)k+\ell}\} \leq \frac{\varepsilon\beta}{\gamma} \left\|\tilde{\theta}_{j-1} - \theta_{\mathrm{S}}\right\|_2$.

By an inductive argument, we can show that for all $i \in [k]$, $\delta_{jk+i} \leq \frac{\varepsilon\beta}{\gamma} \left\|\tilde{\theta}_{j-1} - \theta_{\mathrm{S}}\right\|_2$. It follows that $\left\|\tilde{\theta}_j - \theta_{\mathrm{S}}\right\|_2 \leq \frac{\varepsilon\beta}{\gamma} \left\|\tilde{\theta}_{j-1} - \theta_{\mathrm{S}}\right\|_2$. Therefore, the sequence of classifiers $\tilde{\theta}_1, \tilde{\theta}_2, \ldots$ converges to θ_{S} at a rate of $\left(1 - \varepsilon\frac{\beta}{\gamma}\right)^t$ for $\varepsilon < \frac{\gamma}{\beta}$. It follows that the sequence $\theta_1, \theta_2, \ldots$ converges to θ_{S} at a rate $k\left(1 - \varepsilon\frac{\beta}{\gamma}\right)^t$. □

B SIMULATION

B.1 Additional Details

We simulate a credit score system using the dataset *GiveMeSomeCredit* Kaggle (2011) 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, individuals prefer a *negative* classification. The prediction is based on 11 biographic and financial history features included in the dataset.

In the rest of this section we describe the deterministic "strategic response function" $\mathcal{D}(\theta)$ of the k Groups Respond Slowly model (Example 2) used in the simulation. Recall that the distribution of the group j at round $t \geq j$ is $\mathcal{D}(\theta_{t-j+1})$.

In the simulation, we assume that the world population is finite, consisting of exactly the individuals in the dataset. There are 18,357 individuals (data points). 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 of each group ($d_0^{|j}$) and the "baseline" distribution d_{BL}, from which modifications are made.

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

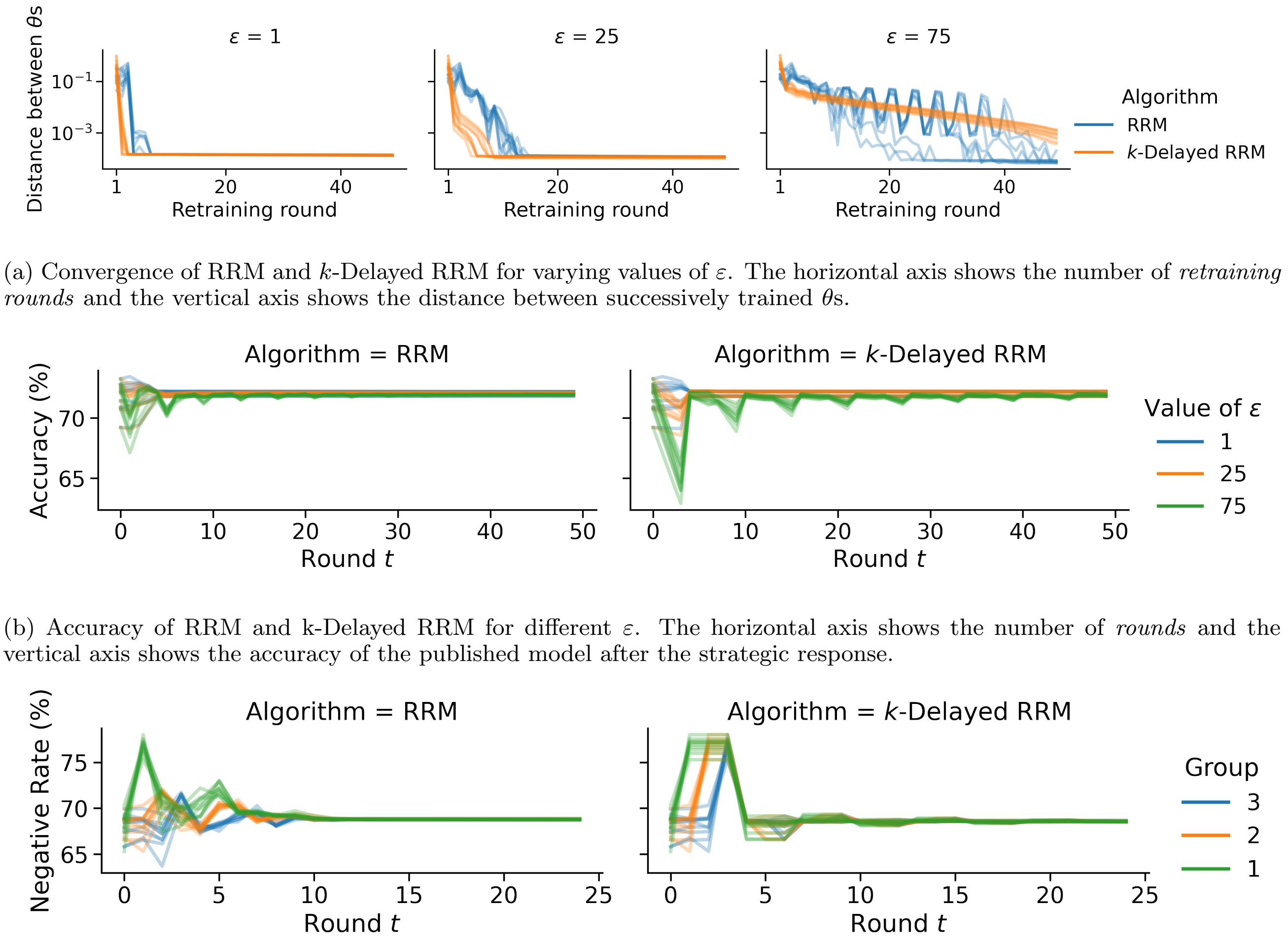
$$x_{\mathrm{BR}}(\theta) \xleftarrow{\arg} \max_{x'} u(x', \theta) - c(x', x),$$

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:

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



(a) Convergence of RRM and k-Delayed RRM for varying values of ε. The horizontal axis shows the number of *retraining rounds* and the vertical axis shows the distance between successively trained θs.

(b) Accuracy of RRM and k-Delayed RRM for different ε. The horizontal axis shows the number of *rounds* and the vertical axis shows the accuracy of the published model after the strategic response.

(c) Negative rate of RRM and k-Delayed RRM for $\varepsilon = 25$. The horizontal axis shows the number of *rounds* and the vertical axis shows the negative rate of the published model after the strategic response.

Figure 4: Comparison of RRM and k-Delayed RRM for for $k = 3$ over three different metrics. We run the experiments from the main text ten times, but with different initial distribution for each group. Each line represents the results of an experiment with a random different group-initialization.

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 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

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

B.2 Different Initial Distributions

We repeat the experiments from Section 4 ten times but with different initial distribution for each group. The initial distribution of group $j \in [k]$, namely $d_0^{|j}$, is generated by sampling *with replacement* $N = 18,357$ individuals from the *GiveMeSomeCredit* dataset. Figure 4 shows that we our findings from the setting where all groups have identical initial distributions also hold for this setup, with the exception of the slower convergence of RRM for some of the initializations when $\varepsilon = 75$