Performative Prediction with Neural Networks

Mehrnaz Mofakhami
Mila, Université de Montréal

Ioannis Mitliagkas
Mila, Université de Montréal
Canada CIFAR AI Chair

Gauthier Gidel
Mila, Université de Montréal
Canada CIFAR AI Chair

Abstract

Performative prediction is a framework for learning models that influence the data they intend to predict. We focus on finding classifiers that are *performatively stable*, i.e. optimal for the data distribution they induce. Standard convergence results for finding a performatively stable classifier with the method of repeated risk minimization assume that the data distribution is Lipschitz continuous to the *model's parameters*. Under this assumption, the loss *must* be strongly convex and smooth in these parameters; otherwise, the method will diverge for some problems. In this work, we instead assume that the data distribution is Lipschitz continuous with respect to the *model's predictions*, a more natural assumption for performative systems. As a result, we are able to significantly relax the assumptions on the loss function. In particular, we do not need to assume convexity with respect to the model's parameters. As an illustration, we introduce a resampling procedure that models realistic distribution shifts and show that it satisfies our assumptions. We support our theory by showing that one can learn performatively stable classifiers with neural networks making predictions about real data that shift according to our proposed procedure.

1 INTRODUCTION

One of the main challenges in many of the decision-making tasks is that the data distribution changes over time. This concept is known as distribution shift or concept drift (Gama et al., 2014; Tsymbal, 2004; Quionero-Candela et al., 2009). With a changing data distribution, the performance of a supervised learning classifier may degrade since it implicitly assumes a static relationship between input and output variables (Fang et al., 2020). *Performative prediction* is a framework introduced by Perdomo et al. (2020) to deal with this problem when the distribution changes as a consequence of model deployment, usually through actions taken based on the model's predictions. For example, election predictions affect campaign activities and, in turn, influence the final election results. This performative behavior arises naturally in many problems of economics, social sciences, and machine learning, such as in loan granting, predictive policing, and recommender systems (Perdomo et al., 2020; Krauth et al., 2022; Ensign et al., 2018).

So far, most works in this area assume strong convexity of the risk function $\theta \mapsto \ell(z; \theta)$, which takes as input the model's parameters θ, and a data point z (Perdomo et al., 2020; Mendler-Dünner et al., 2020; Brown et al., 2022b). For example, Perdomo et al. (2020) show that by assuming strong convexity and smoothness of the loss function along with some regularity assumptions on the data generation process, repeated retraining converges to a performative stable classifier, which is a model that is optimal for the distribution it induces. However, this strong convexity assumption does not hold for most modern ML models, e.g. neural networks. From a different perspective, given a data point (x, y), the risk function can be expressed as a mapping from the prediction $x \mapsto f_\theta(x)$ to a loss between the prediction $\hat{y} := f_\theta(x)$ and the target y, in which case convexity almost always holds. Consider, for example, the Squared Error loss function in a binary classification problem where $\ell(f_\theta(x), y) = \frac{1}{2}(f_\theta(x) - y)^2$ for a data point $z = (x, y)$; it is convex with respect to the model's predictions $f_\theta(x)$, but not necessarily with respect to the model's parameters θ.

With this in mind, we propose a different perspective and formulation that shifts attention from the space of parameters to the space of predictions. More precisely, we require distributions to be functions of the model's prediction function instead of its parameters. The rationale behind this is that in many scenarios with performative effects of a classifier in the loop, the model's predictions are the quantities of interest rather than its parameters. Actually, the framework assumes the data distribution changes as a result of model deployment, and at the time of deployment, it is the final predictions that matter rather than the parameters which led to those predictions.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

Within our formulation, we show that by having a stronger assumption on the distribution map than the original framework, we can relax the convexity condition on the loss function and prove the existence and uniqueness of a performative stable classifier under repeated risk minimization with significantly weaker assumptions with regard to the regularity of the loss. Informally, these assumptions include strong convexity of the loss with respect to the predictions, the boundedness of the derivative of the loss function, and the Lipschitzness of the distribution map with respect to the χ^2 divergence. This more general set of assumptions on the loss function lets us theoretically analyze the performative effects of neural networks with non-convex loss functions; we believe this is a significant step toward bridging the gap between the theoretical performative prediction framework and realistic settings. Our setting and the main theoretical results will be explained in Section 2.

1.1 Background

Before stating our main theoretical contribution, we first recall the key concepts of the performative prediction framework. The main difference between classic supervised learning and the performative prediction framework is that the latter considers the data distribution to be model-dependent, i.e. it assumes that the distribution map directly depends on the model's parameters θ and is denoted by $\theta \mapsto \mathcal{D}(\theta)$. The distribution map is said to satisfy a notion of Lipschitz continuity called *ϵ-sensitivity* if for any θ and θ', $\mathcal{W}_1\left(\mathcal{D}(\theta), \mathcal{D}(\theta')\right) \leq \epsilon\|\theta - \theta'\|_2$, where $\mathcal{W}_1$ denotes the Wasserstein-1 distance. The performance of a model with parameters θ is measured by its *performative risk* under the loss function ℓ which is stated as a function of a data point z and θ:

$$PR(\theta) \overset{\text{def}}{=} \underset{z\sim\mathcal{D}(\theta)}{\mathbb{E}} \ell(z;\theta).$$

A classifier with parameters θ_{PS} is *performatively stable* if it minimizes the risk on the distribution it induces. In other words, it is the fixed point of repeated retraining.

$$\theta_{\text{PS}} = \arg\min_{\theta\in\Theta} \underset{z\sim\mathcal{D}(\theta_{\text{PS}})}{\mathbb{E}} \ell(z;\theta).$$

Perdomo et al. (2020) demonstrate that with an ϵ-sensitive distribution map, γ-strong convexity of ℓ in θ and β-smoothness of ℓ in θ and z are sufficient and *necessary* conditions for repeated risk minimization to converge to a performatively stable classifier if $\frac{\epsilon\beta}{\gamma} < 1$. We show, however, that by slightly changing the assumptions on the distribution map, we can break their negative result regarding the necessary strong convexity and show that one can converge even when the loss is non-convex in θ.

1.2 Our contributions

Our paper provides sufficient conditions for the convergence of repeated risk minimization to a classifier with unique predictions under performative effects in the absence of convexity to the model's parameters. The key idea in our framework is that the distribution map is no longer a function of the parameters θ, but a function of model's predictions, denoted by $\mathcal{D}(f_\theta)$ where f_θ is in $\mathcal{F}$, the set of parameterized functions by $\theta \in \Theta$. We also express the loss ℓ as a function of the prediction $f_\theta(x)$ and the target y, where both can be multi-dimensional. Following is the informal statement of our main theorem.

Theorem 1. *(Informal) If the loss $\ell(f_\theta(x), y)$ is strongly convex in $f_\theta(x)$ with a bounded gradient norm, and the distribution map $f_\theta \mapsto \mathcal{D}(f_\theta)$ is sufficiently Lipschitz with respect to the χ^2 divergence and satisfies a bounded norm ratio condition, then repeated risk minimization converges linearly to a stable classifier with unique predictions.*

We will state this theorem formally in section 2. The critical assumption we make on the distribution map is Lipschitz continuity, which captures the idea that a small change in the model's predictions cannot lead to a large change in the induced data distribution, as measured by the χ^2 divergence. This is more restrictive than the Lipschitz continuity assumption of Perdomo et al. (2020) with $\mathcal{W}_1$, since for the χ^2 divergence to be finite, distributions should have the same support. However, we show that this still holds in realistic settings, and we discuss that this stronger assumption on the distribution map is a price we have to pay to relax the assumptions on the loss function significantly and have convergence guarantees for neural networks with non-convex loss functions.

In section 4, we demonstrate our main results empirically with a *strategic classification* task, which has been used as a benchmark for performative prediction (Perdomo et al., 2020; Miller et al., 2021b; Brown et al., 2022b). Strategic classification involves an institution that deploys a classifier and agents who strategically manipulate their features to alter the classifier's predictions to get better outcomes. We propose a resampling procedure called *Resample-if-Rejected (RIR)* in Section 3 to model the population's strategic responses and show that it results in a distribution map that satisfies the conditions of Theorem 1. Within this process, a sample x is drawn from the base distribution and is rejected with some probability dependent on $f_\theta(x)$ and accepted otherwise; in case of rejection, another sample from the base distribution will be drawn. A real-life example that this procedure may be able to model is regarding posting content on social media. Actually, social media use many ML models to automatically regulate the content posted by the users. Consequently, some users' posts may be rejected by the automatic regulation because their content was considered to violate the platform's community guidelines. In some situations, the authors might consider this rejection unfair and may tweak some parts of the post in order to get accepted. In our experiments, we model this resubmission by a resampling of some strategic features (i.e., features that do not drastically affect the content and that are easily modifiable) of the post.

1.3 Related work

Prior work on performative prediction focused on learning from a data distribution $\mathcal{D}(\theta)$ that could change with the model's parameter θ (Perdomo et al., 2020; Mendler-Dünner et al., 2020; Brown et al., 2022a; Drusvyatskiy and Xiao, 2022; Miller et al., 2021a; Izzo et al., 2021; Maheshwari et al., 2022; Ray et al., 2022; Li and Wai, 2022; Dong and Ratliff, 2021; Jagadeesan et al., 2022). In this work, we propose to strengthen the standard ϵ-sensitivity assumption on the distribution map initially proposed by Perdomo et al. (2020). To a certain extent, we propose a novel ϵ-sensitivity assumption for the performative prediction framework that allows us to relax the convexity assumption on the loss function. Such relaxation is essential if we want to consider the practical setting of classifiers parametrized by neural networks.

At a technical level, our analysis is inspired by Perdomo et al. (2020, Theorem 3.5). However, because our quantity of interest is a distance in the function space (see Theorem 2) our proof significantly differs from Perdomo et al. (2020). We require a different notion of ϵ-sensitivity and an additional assumption (Assumption 2) in order to control the variation of the functional norm defined in Assumption 1.

Various prior works have focused on finding performatively stable classifiers (Perdomo et al., 2020; Brown et al., 2022b; Mendler-Dünner et al., 2020; Li and Wai, 2022), but to the best of our knowledge, none of them analyze the convergence of repeated retraining with loss functions that might be non-convex to the model's parameters.

Exploiting convexity in the model's predictions has previously been explored by Bengio et al. (2005), who noticed that most of the loss functions to train neural networks are convex with respect to the neural network itself. There have been many works trying to leverage this property to show convergence results applied to neural networks in the context of machine learning (Bach, 2017; Chizat and Bach, 2018; Mladenovic et al., 2021). However, none of these results are in the context of performative prediction. Jagadeesan et al. (2022) proposes an algorithm to find classifiers with near-optimal performative risk without assuming convexity. First, their work focuses on a different notion of optimality (namely, performatively optimal points). Second, they focus on regret minimization, while our work is concerned with finding a performatively stable classifier with gradient-based algorithms and having guarantees to make sure we converge to such a stable classifier within a reasonable number of steps.[1] Similarly to our work, Mendler-Dünner et al. (2022) assume that the performativity of a model occurs through its predictions and consider the distribution a function of the predictive model. However, this paper has a different focus entirely; they try to find a set of conditions under which the causal effect of predictions becomes identifiable. Additionally, they focus on a subset of performative prediction problems where predictions only influence the target variable and not the features X. Hence, their analysis does not capture strategic classification.

2 FRAMEWORK AND MAIN RESULTS

To propose our main theorem, we first need to redefine some of the existing concepts. As mentioned earlier, we assume $\mathcal{D}(\theta)$ to be a mapping from the model's prediction function f_θ to a distribution $\mathcal{D}(f_\theta)$ over instances z, where f_θ is in $\mathcal{F}$, the set of parameterized functions by $\theta \in \Theta$. Each instance z is a pair of features and label (x, y). With this new formulation, the objective risk function will be defined as follows:

Definition 2.1 (*Performative Risk*)**.** *Performative risk (PR) is defined as follows:*

$$PR(f_\theta) \overset{def}{=} \mathop{\mathbb{E}}_{z\sim\mathcal{D}(f_\theta)} \ell(f_\theta(x), y).$$

In this work, we focus on finding a performatively stable classifier, which minimizes the risk on the distribution its prediction function entails:

Definition 2.2. *A classifier with parameters θ_{PS} is performatively stable if:*

$$\theta_{PS} = \arg\min_{\theta\in\Theta} \mathop{\mathbb{E}}_{z\sim\mathcal{D}(f_{\theta_{PS}})} \ell(f_\theta(x), y).$$

Repeated retraining is the algorithm we use to find a stable classifier, which is defined formally as follows:

Definition 2.3 (*RRM*)**.** *Repeated Risk Minimization (RRM) refers to the procedure where, starting from an initial θ_0, we perform the following sequence of updates for every $t \geq 0$:*

$$\theta_{t+1} = G(\theta_t) \overset{def}{=} \arg\min_{\theta\in\Theta} \mathop{\mathbb{E}}_{z\sim\mathcal{D}(f_{\theta_t})} \ell(f_\theta(x), y).$$

2.1 Assumptions

In order to provide convergence guarantees for repeated retraining, we require regularity assumptions on the distribution map and the loss function. A natural assumption we make on $\mathcal{D}(.)$ inspired by prior work is Lipschitz continuity, formally referred to as ϵ-*sensitivity*. Intuitively, this assumption states the idea that if two models with similar

[1] For a δ-approximate optimum, Jagadeesan et al. (2022) propose an algorithm that requires $O(1/\delta^d)$ repeated minimizations for the last iterate where d is some notion of dimension. In comparison, in Theorem 2 we require $O(\log(1/\delta))$ minimizations.

prediction functions are deployed, then the induced distributions should also be similar. We refer to the *base distribution* $\mathcal{D}$ as the distribution over (features, label) pairs before any classifier deployment.

Assumption 1. *(A1) [ϵ-sensitivity w.r.t Pearson χ^2 divergence] Suppose the base distribution $\mathcal{D}$ has the probability density function (pdf) p over instances $z = (x, y)$. The distribution map $\mathcal{D}(.)$ which maps f_θ to $\mathcal{D}(f_\theta)$ with the pdf p_{f_θ} is ϵ-sensitive w.r.t Pearson χ^2 divergence, i.e., for all f_θ and $f_{\theta'}$ in $\mathcal{F}$ the following holds:*

$$\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta)) \leq \epsilon \|f_\theta - f_{\theta'}\|^2,$$

where $\|f_\theta - f_{\theta'}\|^2 := \int \|f_\theta(x) - f_{\theta'}(x)\|^2 p(x) dx$ and $\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta)) := \int \frac{\left(p_{f_{\theta'}}(z) - p_{f_\theta}(z)\right)^2}{p_{f_\theta}(z)} dz$

Assumption 2. *(A2) [Bounded norm ratio] The distribution map $\mathcal{D}(.)$ satisfies bounded norm ratio with the parameter $C \geq 1$ if for all $f_\theta, f_{\theta'}, f_{\theta^*} \in \mathcal{F}$:*

$$\|f_\theta - f_{\theta'}\|^2 \leq C \|f_\theta - f_{\theta'}\|^2_{f_{\theta^*}},$$

where

$$\|f_\theta - f_{\theta'}\|^2_{f_{\theta^*}} = \int \|f_\theta(x) - f_{\theta'}(x)\|^2 p_{f_{\theta^*}}(x) dx$$

is a notation for a f_{θ^*}-dependent norm. In other words, this assumption says that

$$\mathbb{E}_{p(x)}[\|f_\theta(x) - f_{\theta'}(x)\|^2] \leq C\, \mathbb{E}_{p_{f_{\theta^*}}(x)}[\|f_\theta(x) - f_{\theta'}(x)\|^2],$$

where $p(x)$ and $p_{f_{\theta^*}}(x)$ are pdfs for the marginal distribution of X according to $\mathcal{D}$ and $\mathcal{D}(f_{\theta^*})$ respectively.

The distribution map satisfies the bounded norm ratio condition if the bounded density ratio property holds, i.e. $p(x) \leq C\ p_{f_\theta}(x)$ for every $f_\theta \in \mathcal{F}$. We will show how the bounded density ratio holds in our example in Section 3.

Our notion of Lipschitz Continuity uses the Pearson χ^2 divergence—interchangeably referred to as χ^2 divergence—to measure the distance between distributions, as opposed to Perdomo et al. (2020) who use $\mathcal{W}_1$ distance. Using χ^2 divergence is more restrictive since the distributions should have the same support for the χ^2 divergence between them to be finite.

As stated in Remark 1, ϵ-sensitivity with respect to χ^2 implies $K\sqrt{\epsilon}$-sensitivity with respect to $\mathcal{W}_1$ for a constant K that depends on the diameter of the space d_{max}. If $\frac{d_{max}}{\sqrt{2}} < \sqrt{\epsilon}$, our notion of ϵ-sensitivity w.r.t χ^2 is indeed stronger than the corresponding notion of ϵ-sensitivity w.r.t $\mathcal{W}_1$. However, in Proposition 1 we show that within our settings, we cannot replace χ^2 with $\mathcal{W}_1$ and still get convergence results.

Remark 1. *For two distributions $\mathcal{D}(x)$ and $\mathcal{D}'(x)$, the $\mathcal{W}_1$ distance is upper bounded by a coefficient of the square root of χ^2 divergence (Peyré et al., 2019, Figure 8.2):*

$$\mathcal{W}_1(\mathcal{D}(x), \mathcal{D}'(x)) \leq \frac{d_{max}}{\sqrt{2}} \sqrt{\chi^2(\mathcal{D}(x), \mathcal{D}'(x))},$$

where $\mathcal{X}$ is a metric space with ground distance d and $d_{max} = \sup_{(x,x')} d(x, x')$ is the diameter of $\mathcal{X}$.
If we define ϵ-sensitivity of $\mathcal{D}(.)$ w.r.t $\mathcal{W}_1$ as

$$\mathcal{W}_1(\mathcal{D}(f_\theta), \mathcal{D}(f_{\theta'})) \leq \epsilon \|f_\theta - f_{\theta'}\|, \qquad (A1)'$$

then ϵ-sensitivity w.r.t χ^2 implies $\frac{d_{max}}{\sqrt{2}}\sqrt{\epsilon}$-sensitivity with respect to $\mathcal{W}_1$:

$$\mathcal{W}_1(\mathcal{D}(f_\theta), \mathcal{D}(f_{\theta'})) \leq \frac{d_{max}}{\sqrt{2}} \sqrt{\epsilon} \|f_\theta - f_{\theta'}\|.$$

Despite the downsides of our assumptions on the distribution map, these assumptions still holds in some realistic settings, an example of which is a resampling procedure proposed in Section 3. The idea of this distribution shift is that individuals are more likely to change their features (by resampling them) if there is a high chance that they will receive an unfavorable classification outcome, which is quantified by the model's prediction.

While imposing a more restrictive assumption on the distribution map, we significantly relax the assumptions on the loss function. In particular, we no longer need to assume that loss is convex to the model's parameters, which opens the door to consider deep neural networks as classifiers in our analysis. We still require some mild assumptions on the loss function ℓ that are as follows:

Assumption 3. *(A3) [Strong convexity w.r.t predictions] The loss function $\ell(f_\theta(x), y)$ which takes as inputs the prediction $f_\theta(x)$ and the target y, is γ-strongly convex in $f_\theta(x)$. More precisely, the following inequality holds for every $f_\theta, f_{\theta'} \in \mathcal{F}$:*

$$\begin{aligned}\ell(f_\theta(x), y) \geq\ & \ell(f_{\theta'}(x), y) + \\ & (f_\theta(x) - f_{\theta'}(x))^\top \nabla_{\hat{y}} \ell(f_{\theta'}(x), y) + \\ & \frac{\gamma}{2} \|f_\theta(x) - f_{\theta'}(x)\|^2,\end{aligned}$$

where $\nabla_{\hat{y}} \ell(f_\theta(x), y)$ is the gradient of the function $\hat{y} \in \mathbb{R}^d \mapsto \ell(\hat{y}, y)$ at $f_\theta(x)$.

Assumption 4. *(A4) [Bounded gradient norm] The loss function $\ell(f_\theta(x), y)$ has bounded gradient norm, i.e., the norm of its gradient with respect to $f_\theta(x)$ is upper bounded with a finite value $M = \sup_{x,y,\theta} \|\nabla_{\hat{y}} \ell(f_\theta(x), y)\|$.*

We can easily see that these two assumptions on ℓ are satisfied by the Squared Error loss: $\ell(f_\theta(x), y) = \frac{1}{2}\|f_\theta(x) - y\|^2$. This function is 1-strongly convex with a bounded gradient norm of $\sqrt{d}$ if y is a one-hot vector in $\mathbb{R}^d$ and $f_\theta(x) \in [0, 1]^d$ for any θ. More broadly, when the predictions are bounded, e.g. in $[0, 1]^d$, then the quantity M in

Assumption 4 always exists for continuously differentiable loss functions which makes it a very mild assumption.

In Section 2.2 we show that if assumptions $A1 - A4$ are satisfied for a distribution map and a loss function, then RRM converges to a unique stable classifier if $\frac{\sqrt{C\epsilon}M}{\gamma}$ is less than 1. Proposition 1 shows that replacing ϵ-sensitivity w.r.t χ^2 (A 1) with ϵ-sensitivity w.r.t $\mathcal{W}_1$ $(A1)'$ while keeping other assumptions would break this convergence result, in the sense that RRM will oscillate between two models forever. This justifies why we cannot use $\mathcal{W}_1$ within our analysis.

Proposition 1. *Suppose that the loss $\ell(f_\theta(x), y)$ is γ-strongly convex in $f_\theta(x)$, and has a derivative bounded by M. If the distribution map satisfies the bounded norm ratio property with a parameter C, and it is ϵ-sensitive w.r.t $\mathcal{W}_1$ $(A1)'$, RRM may diverge for any value of ϵ, particulaly even if $\frac{\sqrt{C\epsilon}M}{\gamma} < 1$.*

Proof. Consider a supervised learning problem where a model with parameters θ uses the prediction function $f_\theta(x) = \frac{\tanh(\theta)+2}{\epsilon}x$ where $x \in [0, 3\epsilon]$. Take the base distribution on X as a uniform distribution over this interval.

The loss function is defined as

$$\ell(f_\theta(x), y) = \frac{-15\gamma}{4}(f_\theta(x)-y)+\frac{\gamma}{2}f_\theta(x)^2+\frac{\gamma}{2}y^2+\gamma(\frac{15}{4})^2.$$

This ℓ is non-negative, γ-strongly convex w.r.t $f_\theta(x)$, and its derivative in $f_\theta(x)$ is bounded.

Let the distribution of X according to $\mathcal{D}(f_\theta)$ be a point mass at $f_\theta(\epsilon^2) = \epsilon(\tanh(\theta) + 2)$ and the distribution of Y be invariant w.r.t f_θ.

$\mathcal{D}(f_\theta)$ is ϵ-sensitive w.r.t the Wasserstein-1 distance:

Choose f_θ and $f_{\theta'}$ arbitrarily. It is easy to see that

$$\mathcal{W}_1(\mathcal{D}(f_\theta), \mathcal{D}(f_{\theta'})) \le \epsilon|\tanh(\theta) - \tanh(\theta')|. \quad (1)$$

$$\begin{aligned}\|f_\theta - f_{\theta'}\|^2 &= \int_0^{3\epsilon} (f_\theta(x) - f_{\theta'}(x))^2 p(x)dx \\ &= \int_0^{3\epsilon} \frac{(\tanh(\theta) - \tanh(\theta'))^2}{\epsilon^2} x^2 p(x)dx \\ &= \frac{(\tanh(\theta) - \tanh(\theta'))^2}{\epsilon^2} \frac{1}{3\epsilon} \int_0^{3\epsilon} x^2 dx \\ &= \frac{(\tanh(\theta) - \tanh(\theta'))^2}{3\epsilon^3} \frac{(3\epsilon)^3}{3} \\ &= 3\left(\tanh(\theta) - \tanh(\theta')\right)^2. \quad (2)\end{aligned}$$

As a result,

$$\|f_\theta - f_{\theta'}\| = \sqrt{3}|\tanh(\theta) - \tanh(\theta')|. \quad (3)$$

Combining (1) and (3) results in the ϵ-sensitivity:

$$\mathcal{W}_1(\mathcal{D}((f_\theta), \mathcal{D}(f_{\theta'}) \le \epsilon\|f_\theta - f_{\theta'}\|.$$

Also, this distribution map satisfies the bounded norm ratio property with any $C > 3$ since:

$$\begin{aligned}&\|f_\theta - f_{\theta'}\|^2_{f_{\theta^*}} \\ &= (f_\theta(\epsilon(\tanh(\theta^*) + 2)) - f_{\theta'}(\epsilon(\tanh(\theta^*) + 2)))^2 \\ &= \left((\tanh(\theta) - \tanh(\theta'))(\tanh(\theta^*) + 2)\right)^2 \\ &> (\tanh(\theta) - \tanh(\theta'))^2, \quad (4)\end{aligned}$$

where we used the fact that $(\tanh(\theta^*) + 2) > 1$.

Putting (2) and (4) together, we can write

$$\|f_\theta - f_{\theta'}\|^2 \le C\,\|f_\theta - f_{\theta'}\|^2_{f_{\theta^*}}$$

for every $C > 3$.

The update rule of RRM is as follows:

$$\begin{aligned}\theta_{t+1} &= \operatorname*{argmin}_\phi \mathbb{E}_{z\sim\mathcal{D}(f_{\theta_t})}[\ell(f_\phi(x), y)] \\ &= \operatorname*{argmin}_\phi \ell(f_\phi(x), y)\Big|_{x=\epsilon(\tanh(\theta_t)+2)}.\end{aligned}$$

Taking the derivative of the loss and setting it to zero results in:

$$(\tanh(\theta_{t+1}) + 2)(\tanh(\theta_t) + 2) = \frac{15}{4}.$$

So if $\theta_t = \tanh^{-1}(\frac{-1}{2})$, then $\theta_{t+1} = \tanh^{-1}(\frac{1}{2})$ and if $\theta_t = \tanh^{-1}(\frac{1}{2})$, then $\theta_{t+1} = \tanh^{-1}(\frac{-1}{2})$.

In conclusion, while the loss function satisfies assumptions (A3) and (A4) and the distribution map satisfies conditions $(A1)'$ and (A2), RRM oscillates between $\tanh^{-1}(\frac{-1}{2})$ and $\tanh^{-1}(\frac{1}{2})$ with $\theta_0 = \tanh^{-1}(\frac{-1}{2})$, for any value of $\epsilon, \gamma, C > 3$, including when $\frac{\sqrt{C\epsilon}M}{\gamma} < 1$. □

2.2 Convergence of RRM

Here we state our main theoretical contribution, which provides sufficient conditions for repeated risk minimization to converge to a stable classifier with unique predictions.

Theorem 2. *Suppose that the loss $\ell(f_\theta(x), y)$ is γ-strongly convex w.r.t $f_\theta(x)$ (A3) and the norm of its gradient w.r.t $f_\theta(x)$ is upper bounded with $M = \sup_{x,y,\theta} \|\nabla_{\hat{y}}\ell(f_\theta(x), y)\|$ (A4). If the distribution map $\mathcal{D}(.)$ is ϵ-sensitive w.r.t Pearson χ^2 divergence (A1) and satisfies bounded norm ratio property with parameter C (A2), then:*

$$\|f_{G(\theta)} - f_{G(\theta')}\| \le \frac{\sqrt{C\epsilon}M}{\gamma}\|f_\theta - f_{\theta'}\|.$$

So if $\frac{\sqrt{C\epsilon}M}{\gamma} < 1$, G is a contractive mapping and RRM converges to a stable classifier at a linear rate:

$$\|f_{\theta_t} - f_{\theta_{PS}}\| \le \alpha,$$

$$\text{for} \quad t \ge (1 - \frac{\sqrt{C\epsilon}M}{\gamma})^{-1}\log(\frac{\|f_{\theta_0} - f_{\theta_{PS}}\|}{\alpha}).$$

As we mentioned earlier, assumptions (A3) and (A4) on ℓ are satisfied by the commonly-used Squared Error loss function, and this holds even in the presence of deep neural networks as predictors. To illustrate our results, we propose the *Resample-if-Rejected* procedure in the following section and show that it satisfies assumptions (A1) and (A2). We provide a proof sketch for Theorem 2 here, though the full proof is available in Supplementary materials.

Proof Sketch. Fix θ and θ' in Θ. Let h and h' be mappings from $\mathcal{F}$ to $\mathbb{R}$ defined as follows:

$$h(f_{\hat{\theta}}) = E_{z\sim\mathcal{D}(f_\theta)}[\ell(f_{\hat{\theta}}(x), y)] = \int \ell(f_{\hat{\theta}}(x), y)p_{f_\theta}(z)dz.$$

$$h'(f_{\hat{\theta}}) = E_{z\sim\mathcal{D}(f_{\theta'})}[\ell(f_{\hat{\theta}}(x), y)] = \int \ell(f_{\hat{\theta}}(x), y)p_{f_{\theta'}}(z)dz.$$

Because of the strong convexity of $\ell(f_\theta(x), y)$ in $f_\theta(x)$ and the fact that $f_{G(\theta)}$ minimizes h, we can show that

$$-\gamma\|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta}^2 \geq \int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}}\ell(f_{G(\theta')}(x), y)p_{f_\theta}(z)dz, \tag{5}$$

where

$$\|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta}^2 = \int \|f_{G(\theta)}(x) - f_{G(\theta')}(x)\|^2 p_{f_\theta}(x)dx.$$

Because of this f_θ-dependent norm, assumption 2 is required so we can remove this dependency later.

Using ϵ-sensitivity of $\mathcal{D}(.)$ w.r.t the χ^2 divergence, and the bounded gradient norm assumption which states that there exists a finite value M such that $M = \sup_{x,y,\theta}\|\nabla_{\hat{y}}\ell(f_\theta(x), y)\|$, alongside the fact that $f_{G(\theta')}$ minimizes h', we derive that

$$\int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}}\ell(f_{G(\theta')}(x), y)p_{f_\theta}(z)dz \geq -M\sqrt{\epsilon}\|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta}\|f_\theta - f_{\theta'}\| \tag{6}$$

which provides a lower bound on the RHS of (5).

Combining (5) and (6) with the fact that the distribution map $\mathcal{D}(.)$ satisfies the bounded norm ratio property with parameter C results in

$$\|f_{G(\theta)} - f_{G(\theta')}\| \leq \frac{\sqrt{C\epsilon}M}{\gamma}\|f_\theta - f_{\theta'}\|.$$

So for $\frac{\sqrt{C\epsilon}M}{\gamma} < 1$, RRM converges to a stable classifier based on the Banach fixed-point theorem.

Setting $\theta = \theta_{t-1}$ and $\theta' = \theta_{PS}$, we can show that this convergence has a linear rate. □

3 ϵ-SENSITIVITY OF THE RIR PROCEDURE

An example of strategic classification, which was introduced in section 1.2, occurs in social media when users' posts get rejected because they violated the platform's policies. In these cases, users usually re-post the same content but with different words in order to get accepted. Inspired by this application, we propose the *Resample-if-Rejected (RIR)* procedure to model distribution shifts. Consider we have a base distribution with pdf p and a function $g : f_\theta(x) \mapsto g(f_\theta(x))$ which indicates the probability of rejection. Here we assume that $f_\theta(x)$ is a scalar. Let RIR(f_θ) be the distribution resulting from deploying a model with prediction function f_θ under this procedure, and take p_{f_θ} as its pdf. The sampling procedure of p_{f_θ} is as follows:

- Take a sample x^* from p.
- Toss a coin whose probability of getting a head is $1 - g(f_\theta(x^*))$. If it comes head, output x^* and if it comes tail, output another sample from p.

Consider X to be a random variable with probability distribution $p(x)$. p_{f_θ} is defined mathematically as

$$\begin{aligned} p_{f_\theta}(x) &= p(x)\Big(1 - g(f_\theta(x))\Big) + p(x)\mathbb{E}_X[g(f_\theta(X))] \\ &= p(x)(1 - g(f_\theta(x)) + C_\theta), \end{aligned}$$

where $C_\theta = \mathbb{E}_X[g(f_\theta(X))] = \int g(f_\theta(x'))p(x')dx'$.

The following theorem shows that the distribution resulting from the RIR procedure satisfies our conditions on the distribution map. This Theorem is proved in the Supplementary materials.

Theorem 3. *If $f_\theta(x) \in [0, 1-\delta]$ $\forall\theta \in \Theta$ for some fixed $0 < \delta < 1$, then for $g(f_\theta(x)) = f_\theta(x) + \delta$, RIR(.) is $\frac{1}{\delta}$-sensitive w.r.t χ^2 divergence (A1) and satisfies the bounded norm ratio property (A2) for $C = \frac{1}{\delta}$.*

Remark 2. *Consider a strategic classification task where the distribution reacts to a model with the prediction function f_θ in accordance with the RIR procedure. Suppose the predictions satisfy $f_\theta(x) \in [0, 1-\delta]$, the label y is in $\{0, 1-\delta\}$, and we use the Squared Error loss $\ell(f_\theta(x), y) = \frac{1}{2}(f_\theta(x) - y)^2$ which is 1-strongly convex. According to Theorem 3, the distribution map is $\frac{1}{\delta}$-sensitive w.r.t χ^2 divergence and satisfies the bounded norm ratio property for $C = \frac{1}{\delta}$. Also, $M = \sup_{x,y,\theta}|\ell'(f_\theta(x), y)|$ is equal to $\sup_{x,y,\theta}|f_\theta(x) - y| = 1 - \delta$. Putting all these together, the convergence rate of RRM in Theorem 2 is equal to:*

$$\frac{\sqrt{C\epsilon}M}{\gamma} = \frac{1-\delta}{\delta}.$$

Hence, whenever we have $\delta > 0.5$, RRM converges to a unique stable classifier.

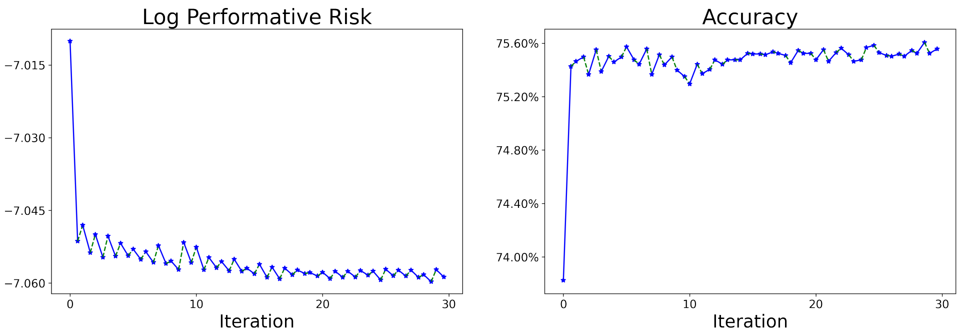

Figure 1: Evolution of log of performative risk (left) and accuracy (right) through iterations of RRM for $\delta = 0.9$. The blue lines show the changes in risk (accuracy) after optimizing on the distribution induced by the last model, and the green lines show the effect of the distribution shift on the risk (accuracy).

We will use this remark in the experiments section.

In supervised learning, x corresponds to a set of features. So far in the RIR procedure, we resample the whole set of features, though we also can resample only a subset of them, and Theorem 3 still holds in this case if strategic and non-strategic features are independent as shown following the proof of this theorem in Supplementary materials. In our simulations in the next section, $x \in \mathbb{R}^d$ is the set of features of individuals applying to get loans from a bank. These features are divided into two sets: *strategic* and *non-strategic*. Strategic features are those that can be (easily) manipulated without affecting the true label, e.g. Number of open credit lines and loans. Non-strategic features, however, can be seen as causes of the label and include monthly income for example. In our experiments, we resample only strategic features as these are the ones that people can manipulate more easily.

4 EXPERIMENTS

We complement our theoretical results with experiments on a credit scoring task and illustrate how they support our claims. We implemented our simulations based on Perdomo et al. (2020)'s code in the Whynot Python package (Miller et al., 2020), and changed it according to our settings so we can use auto-differentiation of PyTorch [2]. The strategic classification task of credit scoring is a two-player game between a bank that predicts the creditworthiness of loan applicants, and individuals who strategically manipulate their features to alter the classification outcome. We run the simulations using Kaggle's *Give Me Some Credit* dataset (Kaggle, 2011), with features $x \in \mathbb{R}^{11}$ corresponding to applicants' information along with their label $y \in \{0, 1\}$, where $y = 1$ indicates that the applicant defaulted and $y = 0$ otherwise.

In our simulations, we assume that the data distribution induced by the classifier f_θ shifts according to the RIR procedure where strategic features are resampled with the probability of rejection $g(f_\theta(x)) = f_\theta(x) + \delta$. Assuming that strategic and non-strategic features are independent, resampling strategic features can be implemented by simply choosing these features from another data point at random. For the classifier, we used a two-layer neural network with a hidden-layer size of 6. The choice of hidden layer size in our network was arbitrary; Figure 3 shows convergence for different hidden size values. In the network, we use a LeakyReLU activation after the first layer, and a scaled-sigmoid activation function after the second layer to bring the outcome $f_\theta(x)$ to the interval $[0, 1-\delta]$. This way we make sure that $g(f_\theta(x)) \in [\delta, 1]$ is a valid probability and the assumption of Theorem 3 is satisfied. Since the outcome $f_\theta(x)$ is in $[0, 1-\delta]$, we change the label 1 to $1-\delta$. So $y = 1-\delta$ corresponds to default, and the higher $f_\theta(x)$ is, the more the chance to get rejected. The objective is to minimize the expectation of the Squared Error loss function over instances, i.e. $\mathbb{E}[\frac{1}{2}(f_\theta(x) - y)^2]$.

The definition of RRM requires solving an exact minimization problem at each optimization step; however, we solve this optimization problem approximately using several steps of gradient descent until the absolute difference of two consecutive risks is less than the tolerance of 10^{-9}. Also, note that running the same configuration twice might result in different plots because of the randomness that exists in the resampling phase.

Figure 1 shows the evolution of the log of performative risk (left) and accuracy (right) through iterations of RRM for $\delta = 0.9$. The blue lines show the changes in risk (accuracy) after optimizing on the distribution induced by the last model, and the green lines show the effect of the distribution shift on the risk (accuracy). We chose to plot the log of performative risk instead its own value only for illustration purposes. As discussed in Remark 2, for this δ, all the conditions of Theorem 2 including $\frac{\sqrt{C\epsilon}M}{\gamma} < 1$ are satisfied, and the

[2] https://github.com/mhrnz/Performative-Prediction-with-Neural-Networks

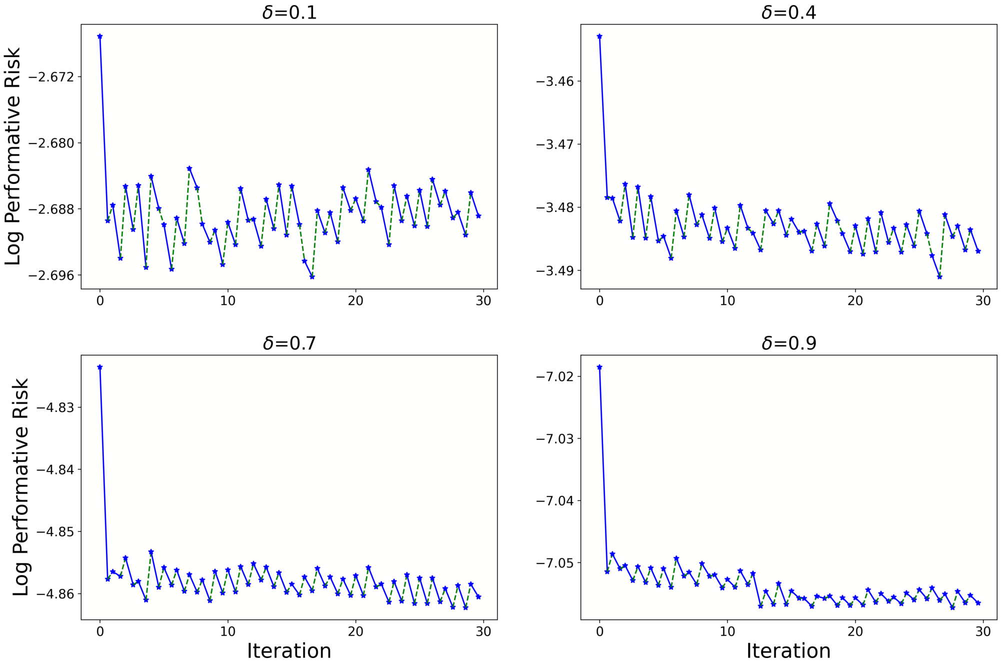

Figure 2: Evolution of log of performative risk for different values of $\delta = 0.1, 0.4, 0.7, 0.9$ through iterations of RRM.

Theorem claims that in this case, RRM converges to a stable model; this is supported by our results in Figure 1.

Figure 2 shows the log of Performative Risk for different values of $\delta = 0.1, 0.4, 0.7, 0.9$. The plot for $\delta = 0.9$ is generated through a different run than Figure 1. Based on Remark 2, for $\delta = 0.7$ and $\delta = 0.9$ we should see convergence behavior, though for $\delta < 0.5$, our theory neither gives a guarantee of convergence nor claims that repeated retraining will diverge, so we might or might not see convergence behavior for $\delta = 0.1$ or $\delta = 0.4$. What we see in Figure 2 is aligned with our expectations. It is important to note that for smaller δ, the value of ϵ which indicates the strength of performative effects is larger, and for high performative effects, it is more difficult for the model to converge since the distribution is allowed to move more after the model's deployment.

On a high level, we interpret the stable classifier to be a model that relies less on non-strategic features for classification. Throughout the training, for a fixed data point $z = (x, y)$ where $x = (x_s, x_f)$ for x_s being the strategic features and x_f being the non-strategic ones, the model sees the same x_f but different values for x_s chosen randomly, all with the same label y. So intuitively, the model would learn to rely less on strategic features and more on non-strategic ones for classification, and this makes it more robust to the strategic behavior of agents.

5 DISCUSSION AND FUTURE WORK

In this paper, we contribute the first set of convergence guarantees for finding performative stable models on problems where the risk is allowed to be non-convex with respect to parameters. This is an important development: our results pertain to modern machine learning models, like neural networks.

We achieve these stronger results by appealing to functional analytical tools, but also making slightly stronger assumptions on the performative feedback loop: rather than assuming that the distribution is ϵ-sensitive to parameters as measured by Wasserstein distance, we instead assume that the distribution is ϵ-sensitive to *predictions* as measured by the χ^2 divergence.

On one hand, only assuming sensitivity to predictions instead of parameters is a step in the right direction. None of the big applications of performative prediction justify sensitivity to model parameters. As a matter of fact, many of the applications would motivate moving one step further in that direction: performative behavior in machine learning systems often manifests as a function of *decisions* or *actions* that rely on a prediction. Those decisions or actions are observed by a population that reacts by changing its behavior. We leave this important problem setting of studying sensitivity to decisions for future work.

On the other hand, χ^2 sensitivity is stronger and implies Wassertstein sensitivity. Furthermore, because we use a

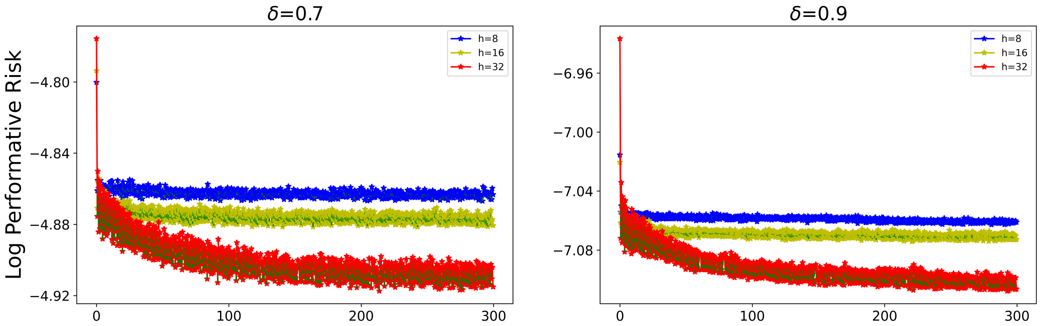

Figure 3: Evolution of log of performative risk through iterations of RRM for different values of hidden size $h = 8, 16, 32$ for $\delta = 0.7$ and $\delta = 0.9$.

variable norm that depends on parameters in our analysis, we make an extra assumption that the fixed norm is upper-bounded by a coefficient of this variable norm. While we provide in Section 3 a well-motivated, concrete example of a performative problem that satisfies both of these conditions, it is nonetheless an interesting open question to wonder how much our analytical assumptions can be loosened.

6 SOCIETAL IMPACT

We believe that the deployment of models that can have an impact on the behavior of people (i.e., are performative) should be considered with care, more especially for some critical applications such as elections or regulation of content on social media platforms. Our work proposes a new analysis of an existing algorithm that aims at learning performatively stable classifiers. Since the nature of our work is mainly theoretical and does not introduce new algorithms, it does not have a direct societal impact beyond the one described in the original paper on performative prediction. However, since our work supports the use of much more powerful models (e.g., NNs) in performative problems, and this increased power comes with increased responsibility, we should be mindful of the potential for undue influence on society while using this framework.

Acknowledgements

We thank Quentin Bertrand, Michael Przystupa, and Reza Bayat for their helpful feedback on the manuscript. We would also like to thank Amjad Almahairi for our insightful discussions.

References

F. Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 2017. (Cited on page 3)

Y. Bengio, N. Roux, P. Vincent, O. Delalleau, and P. Marcotte. Convex neural networks. *Advances in neural information processing systems*, 18, 2005. (Cited on page 3)

G. Brown, S. Hod, and I. Kalemaj. Performative prediction in a stateful world. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022a. (Cited on page 3)

G. Brown, S. Hod, and I. Kalemaj. Performative prediction in a stateful world. *AISTATS*, 2022b. doi: 10.48550/ARXIV.2011.03885. URL https://arxiv.org/abs/2011.03885. (Cited on pages 1, 2, and 3)

L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 2018. (Cited on page 3)

R. Dong and L. J. Ratliff. Approximate regions of attraction in learning with decision-dependent distributions. *arXiv preprint arXiv:2107.00055*, 2021. (Cited on page 3)

D. Drusvyatskiy and L. Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 2022. (Cited on page 3)

D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*, pages 160–171. PMLR, 2018. (Cited on page 1)

T. Fang, N. Lu, G. Niu, and M. Sugiyama. Rethinking importance weighting for deep learning under distribution shift, 2020. URL https://arxiv.org/abs/2006.04662. (Cited on page 1)

J. a. Gama, I. Žliobaitundefined, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46, mar 2014. ISSN 0360-0300. doi: 10.1145/2523813. URL https://doi.org/10.1145/2523813. (Cited on page 1)

Z. Izzo, L. Ying, and J. Zou. How to learn when data reacts to your model: performative gradient descent. In *International Conference on Machine Learning*, pages 4641–4650. PMLR, 2021. (Cited on page 3)

M. Jagadeesan, T. Zrnic, and C. Mendler-Dünner. Regret minimization with performative feedback. *International Conference on Machine Learning*, 2022. (Cited on page 3)

Kaggle. Give me some credit dataset. 2011. URL https://www.kaggle.com/c/GiveMeSomeCredit. (Cited on page 7)

K. Krauth, Y. Wang, and M. I. Jordan. Breaking feedback loops in recommender systems with causal inference, 2022. URL https://arxiv.org/abs/2207.01616. (Cited on page 1)

Q. Li and H.-T. Wai. State dependent performative prediction with stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3164–3186. PMLR, 2022. (Cited on page 3)

D. Liberzon. Calculus of variations and optimal control theory: A concise introduction. *Princeton University Press*, 2012. (Cited on page 11)

C. Maheshwari, C.-Y. Chiu, E. Mazumdar, S. Sastry, and L. Ratliff. Zeroth-order methods for convex-concave min-max problems: Applications to decision-dependent risk minimization. In *International Conference on Artificial Intelligence and Statistics*, pages 6702–6734. PMLR, 2022. (Cited on page 3)

C. Mendler-Dünner, J. Perdomo, T. Zrnic, and M. Hardt. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 2020. (Cited on pages 1 and 3)

C. Mendler-Dünner, F. Ding, and Y. Wang. Anticipating performativity by predicting from predictions. *Advances in Neural Information Processing Systems*, 2022. (Cited on page 3)

J. Miller, C. Hsu, J. Troutman, J. Perdomo, T. Zrnic, L. Liu, Y. Sun, L. Schmidt, and M. Hardt. Whynot, 2020. URL https://doi.org/10.5281/zenodo.3875775. (Cited on page 7)

J. P. Miller, J. C. Perdomo, and T. Zrnic. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, pages 7710–7720. PMLR, 2021a. (Cited on page 3)

J. P. Miller, J. C. Perdomo, and T. Zrnic. Outside the echo chamber: Optimizing the performative risk. *CoRR*, abs/2102.08570, 2021b. URL https://arxiv.org/abs/2102.08570. (Cited on page 2)

A. Mladenovic, I. Sakos, G. Gidel, and G. Piliouras. Generalized natural gradient flows in hidden convex-concave games and gans. In *International Conference on Learning Representations*, 2021. (Cited on page 3)

J. C. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt. Performative prediction. *International Conference on Machine Learning*, 2020. (Cited on pages 1, 2, 3, 4, 7, and 11)

G. Peyré, M. Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. (Cited on page 4)

J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009. ISBN 0262170051. (Cited on page 1)

M. Ray, L. J. Ratliff, D. Drusvyatskiy, and M. Fazel. Decision-dependent risk minimization in geometrically decaying dynamic environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. (Cited on page 3)

A. Tsymbal. The problem of concept drift: Definitions and related work. 05 2004. (Cited on page 1)

Supplementary Materials

A PROOF OF THEOREMS

A.1 Auxiliary Lemmas

In order to prove theorems, we first state a definition and a lemma based on Chapter 1 of Liberzon (2012) that will be used later in the proof of Theorem 2.

In the following, a functional $J : V \mapsto \mathbb{R}$ is a function over a space of functions V.

Definition A.1 (*First variation of a functional*)**.** *Let $J : V \to \mathbb{R}$ be a functional on a function space V, and consider some function $y \in V$. The derivative of J at y, which is called the first variation (a.k.a Gateaux derivative), is also a functional on V and is defined as follows:*
A linear functional $\delta J|_y : V \to \mathbb{R}$ is called the first variation of J at y if for all η and all α we have

$$J(y+\alpha\eta) = J(y) + \delta J|_y(\eta)\alpha + o(\alpha), \tag{7}$$

where $\lim_{\alpha\to 0} \frac{o(\alpha)}{\alpha} = 0$. In other words:

$$\delta J|_y(\eta) = \lim_{\alpha\to 0} \frac{J(y+\alpha\eta) - J(y)}{\alpha}. \tag{8}$$

Lemma 1 (*First-order necessary condition for optimality in a constrained function space*)**.** *If y is a minimizer of J, then for every $\eta \in V$ such that $y + \alpha\eta \in V$, $\forall\alpha \in [0, \delta]$ for some $\delta > 0$: $\delta J|_y(\eta) \geq 0$.*

Proof of Lemma 1. Let η be an element of V such that $y + \alpha\eta \in V$, $\forall\alpha \in [0, \delta]$ for some $\delta > 0$ and let $g(\alpha) := J(y + \alpha\eta)$ with domain $[0, \delta]$. From (8) we can conclude that $\delta J|_y(\eta) = g'(0)$, so it suffices to show that $g'(0) \geq 0$.
Since y is a minimizer of J, then 0 is a minimizer of g. The first-order Taylor approximation of g around 0 is as follows:

$$g(\alpha) = g(0) + g'(0)\alpha + o(\alpha), \tag{9}$$

where $\lim_{\alpha\to 0} \frac{o(\alpha)}{\alpha} = 0$. Now I want to show that $g'(0) \geq 0$. Suppose that $g'(0) < 0$. Then there exists an $\epsilon > 0$ small enough such that for any $\alpha < \epsilon$, $|\frac{o(\alpha)}{\alpha}| < |g'(0)|$, i.e. $|o(\alpha)| < |g'(0)\alpha|$. Therefore, for $\alpha < \epsilon$ we can write the following inequality using (9):

$$g(\alpha) - g(0) < g'(0)\alpha + |g'(0)\alpha|. \tag{10}$$

Since we assumed $g'(0) < 0$ and $\alpha > 0$, (10) will result in $g(\alpha) - g(0) < 0$, which contradicts the fact that g is minimum at 0. This gives us the proof that $g'(0) \geq 0$, hence $\delta J|_y(\eta) \geq 0$. □

A.2 Proof of Theorem 2

For proving Theorem 2, we were inspired by the proof of Theorem 3.5 in Perdomo et al. (2020), but our proof is significantly different from theirs since our analysis is dependent on the prediction function and we need to use infinite-dimensional optimization.

Proof. Fix θ and θ' in Θ. Let $h : \mathcal{F} \mapsto \mathbb{R}$ and $h' : \mathcal{F} \mapsto \mathbb{R}$ be two functionals defined as follows:

$$h(f_{\hat{\theta}}) = E_{z\sim\mathcal{D}(f_\theta)}[\ell(f_{\hat{\theta}}(x), y)] = \int \ell(f_{\hat{\theta}}(x), y) p_{f_\theta}(z) dz \tag{11}$$

$$h'(f_{\hat{\theta}}) = E_{z\sim\mathcal{D}(f_{\theta'})}[\ell(f_{\hat{\theta}}(x), y)] = \int \ell(f_{\hat{\theta}}(x), y) p_{f_{\theta'}}(z) dz \tag{12}$$

where each data point z is a pair of features x and label y.

For a fixed $z = (x, y)$, due to strong convexity of $\ell(f_\theta(x), y)$ in $f_\theta(x)$ we have:

$$\ell(f_{G(\theta)}(x), y) - \ell(f_{G(\theta')}(x), y) \geq \big(f_{G(\theta)}(x) - f_{G(\theta')}(x)\big)^\top \nabla_{\hat{y}} \ell(f_{G(\theta')}(x), y) + \frac{\gamma}{2} \|f_{G(\theta)}(x) - f_{G(\theta')}(x)\|^2. \tag{13}$$

Now take integral over z, and define $\|f_{G(\theta)} - f_{G(\theta')}\|^2_{f_\theta} = \int \|f_{G(\theta)}(x) - f_{G(\theta')}(x)\|^2 p_{f_\theta}(z) dz$ $\big($which is equal to $\int \|f_{G(\theta)}(x) - f_{G(\theta')}(x)\|^2 p_{f_\theta}(x) dx\big)$:

$$h(f_{G(\theta)}) - h(f_{G(\theta')}) \geq \left(\int \big(f_{G(\theta)}(x) - f_{G(\theta')}(x)\big)^\top \nabla_{\hat{y}} \ell(f_{G(\theta')}(x), y) p_{f_\theta}(z) dz\right) + \frac{\gamma}{2} \|f_{G(\theta)} - f_{G(\theta')}\|^2_{f_\theta}. \tag{14}$$

Similarly:

$$h(f_{G(\theta')}) - h(f_{G(\theta)}) \geq \left(\int \big(f_{G(\theta')}(x) - f_{G(\theta)}(x)\big)^\top \nabla_{\hat{y}} \ell(f_{G(\theta)}(x), y) p_{f_\theta}(z) dz\right) + \frac{\gamma}{2} \|f_{G(\theta)} - f_{G(\theta')}\|^2_{f_\theta}. \tag{15}$$

Knowing that $f_{G(\theta)}$ minimizes h, it is enough to show that

$$\int \big(f_{G(\theta')}(x) - f_{G(\theta)}(x)\big)^\top \nabla_{\hat{y}} \ell(f_{G(\theta)}(x), y) p_{f_\theta}(z) dz \geq 0. \tag{16}$$

to conclude:

$$-\gamma \|f_{G(\theta)} - f_{G(\theta')}\|^2_{f_\theta} \geq \int \big(f_{G(\theta)}(x) - f_{G(\theta')}(x)\big)^\top \nabla_{\hat{y}} \ell(f_{G(\theta')}(x), y) p_{f_\theta}(z) dz. \tag{17}$$

This is a key inequality that we will use later in the proof.

Now let's prove inequality (16) using lemma 1. Let $\eta = f_{G(\theta')} - f_{G(\theta)}$. For every $\alpha \in [0, 1]$, $f_{G(\theta)} + \alpha\eta$ is in the function space (supposing it is convex). We know that $f_{G(\theta)}$ is a minimizer of h, so using Lemma 1, $\delta h|_{f_{G(\theta)}}(\eta) \geq 0$. We can write $\delta h|_{f_{G(\theta)}}(\eta)$ as follows:

$$\begin{aligned}
\delta h|_{f_{G(\theta)}}(\eta) &= \lim_{\alpha \to 0} \frac{h(f_{G(\theta)} + \alpha\eta) - h(f_{G(\theta)})}{\alpha} \\
&= \lim_{\alpha \to 0} \int \frac{\ell(f_{G(\theta)}(x) + \alpha\eta(x), y) - \ell(f_{G(\theta)}(x), y)}{\alpha} p_{f_\theta}(z) dz \\
&= \int \lim_{\alpha \to 0} \frac{\ell(f_{G(\theta)}(x) + \alpha\eta(x), y) - \ell(f_{G(\theta)}(x), y)}{\alpha} p_{f_\theta}(z) dz \\
&= \int \eta(x)^\top \nabla_{\hat{y}} \ell(f_{G(\theta)}(x), y) p_{f_\theta}(z) dz \\
&= \int (f_{G(\theta')}(x) - f_{G(\theta)}(x))^\top \nabla_{\hat{y}} \ell(f_{G(\theta)}(x), y) p_{f_\theta}(z) dz.
\end{aligned} \tag{18}$$

Knowing $\delta h|_{f_{G(\theta)}}(\eta) \geq 0$ completes the proof of (16).

Now recall that there exists M such that $M = \sup_{x,y,\theta} \|\nabla_{\hat{y}} \ell(f_\theta(x), y)\|$ and the distribution map over data is ϵ-sensitive w.r.t Pearson χ^2 divergence, i.e.

$$\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta)) \leq \epsilon \|f_\theta - f_{\theta'}\|^2. \tag{19}$$

With this in mind, we do the following calculations:

$$\begin{aligned}
&\left|\int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}}\ell(f_{G(\theta')}(x), y) p_{f_\theta}(z) dz - \int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}}\ell(f_{G(\theta')}(x), y) p_{f_{\theta'}}(z) dz\right| \\
&= \left|\int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}}\ell(f_{G(\theta')}(x), y) \left(p_{f_\theta}(z) - p_{f_{\theta'}}(z)\right) dz\right| \\
&\overset{(*)}{\leq} \int \left|\left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}}\ell(f_{G(\theta')}(x), y) \left(p_{f_\theta}(z) - p_{f_{\theta'}}(z)\right)\right| dz \\
&\leq M \int \left|\|f_{G(\theta)}(x) - f_{G(\theta')}(x)\| \left(p_{f_\theta}(z) - p_{f_{\theta'}}(z)\right)\right| dz \\
&= M \int \left|\|f_{G(\theta)}(x) - f_{G(\theta')}(x)\| \frac{p_{f_\theta}(z) - p_{f_{\theta'}}(z)}{p_{f_\theta}(z)} p_{f_\theta}(z)\right| dz \\
&= M \left|\int |\|f_{G(\theta)}(x) - f_{G(\theta')}(x)\| \frac{p_{f_\theta}(z) - p_{f_{\theta'}}(z)}{p_{f_\theta}(z)} | p_{f_\theta}(z) dz\right| \\
&\overset{\text{Cauchy-Schwarz Ineq.}}{\leq} M \left(\int \|f_{G(\theta)}(x) - f_{G(\theta')}(x)\|^2 p_{f_\theta}(z) dz\right)^{\frac{1}{2}} \left(\int \left(\frac{p_{f_\theta}(z) - p_{f_{\theta'}}(z)}{p_{f_\theta}(z)}\right)^2 p_{f_\theta}(z) dz\right)^{\frac{1}{2}} \\
&= M \|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta} \sqrt{\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta))} \\
&\overset{(19)}{\leq} M\sqrt{\epsilon} \|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta} \|f_\theta - f_{\theta'}\|
\end{aligned}$$

$(*)$ comes from the fact that $\left|\int f(x)dx\right| \leq \int |f(x)|dx$, and the Cauchy-Schwarz inequality states that $|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$.

What we can conclude from the above derivations is that:

$$\begin{aligned}
&\left|\int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}}\ell(f_{G(\theta')}(x), y) p_{f_\theta}(z) dz - \int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}}\ell(f_{G(\theta')}(x), y) p_{f_{\theta'}}(z) dz\right| \\
&\leq M\sqrt{\epsilon} \|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta} \|f_\theta - f_{\theta'}\|. \qquad (20)
\end{aligned}$$

Similar to inequality (16), since $f_{G(\theta')}$ minimizes h', one can prove:

$$\int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}}\ell(f_{G(\theta')}(x), y) p_{f_{\theta'}}(z) dz \geq 0. \qquad (21)$$

From (17) we know that $\int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}}\ell(f_{G(\theta')}(x), y) p_{f_\theta}(z) dz$ is negative, so with this fact alongside (20) and (21), we can write:

$$\int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^\top \nabla_{\hat{y}}\ell(f_{G(\theta')}(x), y) p_{f_\theta}(z) dz \geq -M\sqrt{\epsilon} \|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta} \|f_\theta - f_{\theta'}\|. \qquad (22)$$

Combining (17) and (22), we will get:

$$\begin{aligned}
&\gamma \|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta}^2 \leq M\sqrt{\epsilon} \|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta} \|f_\theta - f_{\theta'}\| \\
&\Rightarrow \|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta} \leq \frac{\sqrt{\epsilon} M}{\gamma} \|f_\theta - f_{\theta'}\| \qquad (23)
\end{aligned}$$

Since the distribution map satisfies the bounded norm ratio assumption with parameter C, we can write:

$$\begin{aligned}
\|f_{G(\theta)} - f_{G(\theta')}\|^2 &= \int \|f_{G(\theta)}(x) - f_{G(\theta')}(x)\|^2 p(x) dx \\
&\leq C \int \|f_{G(\theta)}(x) - f_{G(\theta')}(x)\|^2 p_{f_\theta}(x) dx \\
&= C \|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta}^2 \qquad (24)
\end{aligned}$$

Consequently,

$$\|f_{G(\theta)} - f_{G(\theta')}\| \leq \sqrt{C} \|f_{G(\theta)} - f_{G(\theta')}\|_{f_\theta} \qquad (25)$$

Using (25) in (23) results in:

$$\|f_{G(\theta)} - f_{G(\theta')}\| \leq \frac{\sqrt{C\epsilon}M}{\gamma}\|f_\theta - f_{\theta'}\|. \tag{26}$$

So if $\frac{\sqrt{C\epsilon}M}{\gamma} < 1$, G is a contractive mapping and RRM converges to a stable classifier based on Banach fixed point theorem.

If we set $\theta = \theta_{t-1}$ and $\theta' = \theta_{PS}$ for θ_{PS} being a stable classifier, we know that $G(\theta) = \theta_t$ and $G(\theta') = \theta_{PS}$. So we will have:

$$\begin{aligned}\|f_{\theta_t} - f_{\theta_{PS}}\| &\leq \frac{\sqrt{C\epsilon}M}{\gamma}\|f_{\theta_{t-1}} - f_{\theta_{PS}}\| \\ &\leq (\frac{\sqrt{C\epsilon}M}{\gamma})^t\|f_{\theta_0} - f_{\theta_{PS}}\|\end{aligned} \tag{27}$$

We can easily see that for $t \geq (1 - \frac{\sqrt{C\epsilon}M}{\gamma})^{-1}\log(\frac{\|f_{\theta_0} - f_{\theta_{PS}}\|}{\alpha})$,

$$(\frac{\sqrt{C\epsilon}M}{\gamma})^t\|f_{\theta_0} - f_{\theta_{PS}}\| \leq \alpha$$

So based on (27),

$$\|f_{\theta_t} - f_{\theta_{PS}}\| \leq \alpha.$$

which shows that RRM converges to a stable classifier at a linear rate. □

A.3 Proof of Theorem 3

Proof. As explained in section 3 of the paper, the pdf of $p_{f_\theta}(x)$ is as follows:

$$\begin{aligned}p_{f_\theta}(x) &= p(x)\Big(1 - g(f_\theta(x))\Big) + p(x)\mathbb{E}_X[g(f_\theta(X))] \\ &= p(x)(1 - g(f_\theta(x)) + C_\theta).\end{aligned} \tag{28}$$

where $C_\theta = \mathbb{E}_X[g(f_\theta(X))] = \int p(x')g(f_\theta(x'))dx'$.

For $g(f_\theta(x)) = f_\theta(x) + \delta$, we have:

$$p_{f_\theta}(x) = p(x)(1 - f_\theta(x) - \delta + C_\theta). \tag{29}$$

where $\delta \leq C_\theta \leq 1$ since $0 \leq f_\theta(x) \leq 1 - \delta$, so $\delta \leq g(f_\theta(x)) \leq 1$ for every x.

In the RIR procedure, the distribution of the label y given x is not affected by the predictions so for every $z = (x, y)$ we have $p_{f_\theta}(z) = p_{f_\theta}(x)p(y|x)$ for any f_θ. This results in the following equality:

$$\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta)) = \int \frac{(p_{f_{\theta'}}(z) - p_{f_\theta}(z))^2}{p_{f_\theta}(z)}dz = \int \frac{(p_{f_{\theta'}}(x) - p_{f_\theta}(x))^2}{p_{f_\theta}(x)}dx$$

Now we can prove that this distribution map is ϵ-sensitive w.r.t χ^2 divergence for $\epsilon = \frac{1}{\delta}$:

$$\begin{aligned}
\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta)) &= \int \frac{(p_{f_{\theta'}}(x) - p_{f_\theta}(x))^2}{p_{f_\theta}(x)} dx \\
&= \int \frac{p(x)^2 (f_\theta(x) - f_{\theta'}(x) - (C_\theta - C_{\theta'}))^2}{p(x)(1 - f_\theta(x) - \delta + C_\theta)} dx \\
&\overset{C_\theta \geq \delta}{\leq} \frac{1}{\delta} \int p(x) \Big[(f_\theta(x) - f_{\theta'}(x))^2 + (C_\theta - C_{\theta'})^2 - 2(f_\theta(x) - f_{\theta'}(x))(C_\theta - C_{\theta'}) \Big] dx \\
&= \frac{1}{\delta} \bigg[\Big(\int p(x)(f_\theta(x) - f_{\theta'}(x))^2 dx \Big) + (C_\theta - C_{\theta'})^2 - 2(C_\theta - C_{\theta'}) \int p(x)(f_\theta(x) - f_{\theta'}(x)) dx \bigg] \\
&\overset{(*')}{=} \frac{1}{\delta} \bigg[\Big(\int p(x)(f_\theta(x) - f_{\theta'}(x))^2 dx \Big) + (C_\theta - C_{\theta'})^2 - 2(C_\theta - C_{\theta'})^2 \bigg] \\
&= \frac{1}{\delta} \bigg[\int p(x)(f_\theta(x) - f_{\theta'}(x))^2 dx - (C_\theta - C_{\theta'})^2 \bigg] \\
&\leq \frac{1}{\delta} \int p(x)(f_\theta(x) - f_{\theta'}(x))^2 dx \\
&= \frac{1}{\delta} \| f_\theta - f_{\theta'} \|^2
\end{aligned} \tag{30}$$

Where $(*')$ comes from the fact that $\int p(x)(f_\theta(x) - f_{\theta'}(x))dx = C_\theta - C_{\theta'}$.

Since $\delta \leq C_\theta \leq 1 \; \forall \theta$, it is easy to see that for any f_{θ^*} and for any x:

$$\frac{p(x)}{p_{f_{\theta^*}}(x)} = \frac{1}{1 - g(f_{\theta^*}(x)) + C_{\theta^*}} \leq \frac{1}{\delta}. \tag{31}$$

Consequently,

$$\mathbb{E}_p[(f_\theta - f_{\theta'})^2] \leq \frac{1}{\delta} \, \mathbb{E}_{p_{f_{\theta^*}}}[(f_\theta - f_{\theta'})^2].$$

So the distribution map satisfies the bounded ratio condition for $C = \frac{1}{\delta}$.

The case where we only resamply strategic features. Suppose that features x are divided into strategic features x_s and non-strategic features x_f, i.e. $x = (x_s, x_f)$, and we resample only strategic features with probability $g(f_\theta(x))$ which is the probability of rejection. The pdf of p_{f_θ} would be as follows:

$$p_{f_\theta}(x) = p(x)(1 - g(f_\theta(x))) + \int_{x'} p(x'_s, x'_f = x_f) \; g(f_\theta(x')) \; p(x_s | x_f) dx' \tag{32}$$

Since we only resample strategic features, the integral should be taken over those samples that have the same non-strategic features as x.

Assuming that strategic and non-strategic features are independent, we can re-write (32) as follows:

$$\begin{aligned}
p_{f_\theta}(x) =& p(x)(1 - g(f_\theta(x))) + \int_{x'} p(x'_s, x'_f = x_f) \; g(f_\theta(x')) \; p(x_s | x_f) dx' \\
&= p(x)(1 - g(f_\theta(x))) + \int_{x'} g(f_\theta(x')) p_{X_s}(x'_s) p_{X_f}(x_f) p_{X_s}(x_s) dx' \\
&= p(x)(1 - g(f_\theta(x))) + \int_{x'} g(f_\theta(x')) p_{X_s}(x'_s) p(x) dx' \\
&= p(x) \Big((1 - g(f_\theta(x))) + \int_{x'} g(f_\theta(x')) p_{X_s}(x'_s) dx' \Big)
\end{aligned} \tag{33}$$

where p_{X_s} and p_{X_f} refer to the marginal distributions of srategic and non-strategic features respectively.

Taking $C'_\theta = \int_{x'} g(f_\theta(x')) p_{X_s}(x'_s) dx'$, $p_{f_\theta}(x) = p(x) \Big(1 - g(f_\theta(x)) + C'_\theta \Big)$ has the same form as (28) with C_θ replaced with C'_θ, so the given proof of Theorem 3 applies to this case as well, hence Theorem 3 holds for this case. □