Performative Prediction with Neural Networks

Mehrnaz Mofakhami
Mila, Université de Montréal
mehrnaz.mofakhami@mila.quebec

Ioannis Mitliagkas
Mila, Université de Montréal
ioannis@mila.quebec

Gauthier Gidel
Mila, Université de Montréal
gidelgau@mila.quebec

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 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

Performative prediction is a framework introduced by [22] to deal with the problem of distribution shift or concept drift ([9, 25, 23]) when the distribution changes as a consequence of the model's 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. Other natural examples in economics, social sciences, and Machine Learning include loan granting, predictive policing, and recommender systems [22, 13, 8].

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 datapoint z ([22, 17, 3]). However, this strong convexity assumption does not hold for most modern ML models, e.g. neural networks. From a different perspective, given a datapoint $z = (x, y)$, 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. For example, the Squared Error loss function $\ell(f_\theta(z), z) = (f_\theta(x) - y)^2$ is convex with respect to $f_\theta(x)$, but not necessarily with respect to θ.

With this in mind, we propose a formulation that shifts attention from the space of parameters to the space of predictions and require distributions to be functions of the *model's prediction function*. We believe this is a more natural assumption, since the framework assumes the data distribution changes as a result of model's deployment, and at the time of deployment, it is the final predictions that have performative effects rather than the parameters. Within our formulation, we show that by having a

Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

slightly 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. This more general set of assumptions on the loss function lets us analyze theoretically 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.

Background Before stating our main theoretical contribution, we first need to recall the key concepts of the performative prediction framework. This framework 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, defined as $PR(\theta) \stackrel{\text{def}}{=} \mathbb{E}_{z\sim\mathcal{D}(\theta)}\,\ell(z;\theta)$. A classifier with parameters θ_{PS} is *performatively stable* if it minimizes the risk on the distribution it induces: $\theta_{\text{PS}} = \arg\min_\theta \mathbb{E}_{z\sim\mathcal{D}(\theta_{\text{PS}})}\,\ell(z;\theta)$. Perdomo et. al in [22] 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. We show, however, that by slightly changing the assumptions on the distribution map, convexity in θ is no longer a necessary condition to have convergence guarantees.

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 is that the distribution map is no longer a function of the parameters θ, but a function of model's prediction function f_θ, denoted by $\mathcal{D}(f_\theta)$. We also express the loss ℓ as a function of the prediction $f_\theta(x)$ and the target y. 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 derivative, 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 important 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 [22] 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 believe 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 ([22, 19, 3]). 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 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.

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 [22, 17, 4, 7, 20, 10, 16, 24, 14, 6, 11]. In this work, we propose to strengthen the standard ϵ-sensitivity assumption on the distribution map initially proposed by [22]. 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.

Exploiting convexity in model's predictions has previously been explored by [2] 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 [1, 5, 21]. However, none of these results

are in the context of performative prediction. Jagadeesan et al. [11] 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]

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 $\mathcal{D}(.)$ 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). In this new formulation, the objective risk function, *Performative Risk (PR)*, is defined as

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

In this work, we focus on finding a performatively stable classifier with parameters θ_{PS}, which minimizes the risk on the distribution its prediction function entails:

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

To find such a stable classifier, we use *Repeated Risk Minimization (RRM)* which refers to the procedure where, starting from an initial θ_0, we perform the following sequence of updates for every $t \geq 0$:

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

We can solve this minimization problem in practice by using standard training methods such as gradient descent. Repeating this process corresponds to retraining on the most recent data.

Assumptions In order to provide convergence guarantees for repeated retraining, we need some kind of 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 prediction functions are deployed, then the induced distributions should also be similar. We use Pearson χ^2 divergence—interchangeably referred to as χ^2 divergence—to measure the distance between distributions.

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

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

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

where $\|f_\theta - f_{\theta'}\|^2_{\theta^*} = \int (f_\theta(x) - f_{\theta'}(x))^2 p_{f_{\theta^*}}(z)dz$ is a notation for a θ^*-dependent norm. In other words, this assumption says that

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

where $p(z)$ is the pdf of the base distribution $\mathcal{D}$, and p_{f_θ} is the pdf of the distribution induced by f_θ.

[1] For a δ-approximate optimum, [11] 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.

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 $f_\theta \in \mathcal{F}$. The bounded density ratio property holds in our example in Section 3.

The followings are the two assumptions we make on the loss function ℓ:

A3 (*Strong convexity w.r.t predictions*) A 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 $f_\theta(x)$ if the following inequality holds for every $f_\theta, f_{\theta'} \in \mathcal{F}$:

$$\ell(f_\theta(x), y) \geq \ell(f_{\theta'}(x), y) + \ell'(f_{\theta'}(x), y)\,(f_\theta(x) - f_{\theta'}(x)) + \frac{\gamma}{2}|f_\theta(x) - f_{\theta'}(x)|^2.$$

A4 (*Bounded derivative*) A loss function $\ell(f_\theta(x), y)$ has bounded derivative if its derivative with respect to $f_\theta(x)$ is upper bounded with a finite value $M = \sup_{x,y,\theta} |\ell'(f_\theta(x), y)|$.

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 in $f_\theta(x)$ with a derivative bounded by 1 if $y \in \{0, 1\}$ and $f_\theta(x) \in [0, 1]$ for any θ.

Convergence of RRM Now we can state our main theorem 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 its derivative w.r.t $f_\theta(x)$ is bounded with $M = \sup_{x,y,\theta} |\ell'(f_\theta(x), y)|$ (A4). If the distribution map $\mathcal{D}(.)$ is ϵ-sensitive w.r.t Pearson χ^2 divergence (A1) and satisfies bounded norm ratio propoerty with parameter C (A2), then:*

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

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:

$$\|f_{\theta_t} - f_{\theta_{PS}}\| \leq \alpha,$$
$$\textit{for } t \geq (1 - \frac{\sqrt{C\epsilon}M}{\gamma})^{-1} \log(\frac{\|f_{\theta_0} - f_{\theta_{PS}}\|}{\alpha}).$$

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). The proof of Theorem 2 is available in the Supplementary Material section.

3 ϵ-sensitivity of the RIR procedure

An example of strategic classification, which was introduced in Section 1, 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. Let $\text{RIR}(f_\theta)$ be the distribution resulted from deploying a model with prediction function f_θ under this procedure, and take p_{f_θ} as its pdf. To sample from p_{f_θ}, we first take a sample x^* from p, and then we toss a coin whose probability of getting a head is $1 - g(f_\theta(x))$. If it comes head, we output x^* and if it comes tail, we output another sample from p.

p_{f_θ} is defined mathematically as $p_{f_\theta}(x) = p(x)\Big(1 - g(f_\theta(x))\Big) + p(x)\mathbb{E}_X[g(f_\theta(X))]$

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 Material section.

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 $C = \frac{1}{\delta}$.*

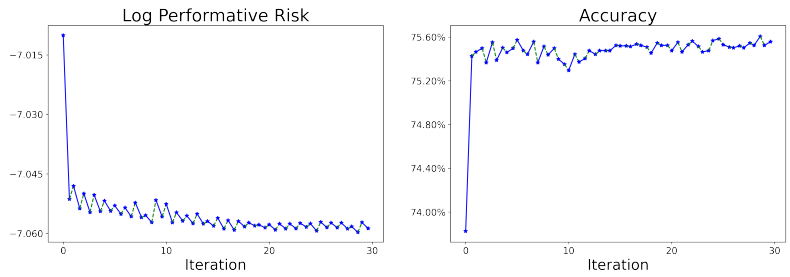

Figure 1: Evolution of log of performative risk (left) and accuracy (right) through iterations of RRM for $\delta = 0.9$. The blue lines show the changes in risk (accuracy) after optimizing on the distribution induced by the last model, and the green lines show the effect of the distribution shift on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

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 the code of [22] in the Whynot Python package [18], and changed it according to our settings so we can use auto-differentiation of PyTorch. This strategic classification task 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 ([12]), which consists of features $x \in \mathbb{R}^d$ corresponding to applicants' information along with their label $y \in \{0, 1\}$, where $y = 1$ indicates that the applicant defaulted and $y = 0$ otherwise.

Individual 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 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 $g(f_\theta(x)) = f_\theta(x) + \delta$. For the classifier, we use a two-layer neural network with a scaled-sigmoid activation function after the second layer to bring the outcome $f_\theta(x)$ to the interval $[0, 1 - \delta]$ to 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 $1 - \delta$. 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}.

Figure 1 shows the evolution of log of performative risk (left) and accuracy (right) through iterations of RRM for $\delta = 0.9$. For this δ, all the conditions of Theorem 2 including $\frac{\sqrt{C\epsilon}M}{\gamma} < 1$ are satisfied, and the Theorem claims that in this case, RRM converges to a stable model; this is supported by our results in Figure 1. Additional experiments on the effect of δ is provided in the Supplementary Material section.

5 Conclusion

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-1 distance, we instead assume that the distribution is ϵ-sensitive to *predictions* as measured by the χ^2 divergence. While we provide in Section 3 a well-motivated, concrete example of a performative problem that satisfies our proposed conditions on the distribution map, it is nonetheless an interesting open question how much our analytical assumptions can be loosened.

References

[1] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 2017.

[2] Yoshua Bengio, Nicolas Roux, Pascal Vincent, Olivier Delalleau, and Patrice Marcotte. Convex neural networks. *Advances in neural information processing systems*, 18, 2005.

[3] Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. *AISTATS*, 2022.

[4] Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022.

[5] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 2018.

[6] Roy Dong and Lillian J Ratliff. Approximate regions of attraction in learning with decision-dependent distributions. *arXiv preprint arXiv:2107.00055*, 2021.

[7] Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 2022.

[8] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*, pages 160–171. PMLR, 2018.

[9] João Gama, Indrundefined Žliobaitundefined, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46, mar 2014.

[10] Zachary Izzo, Lexing Ying, and James Zou. How to learn when data reacts to your model: performative gradient descent. In *International Conference on Machine Learning*, pages 4641–4650. PMLR, 2021.

[11] Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. Regret minimization with performative feedback. *International Conference on Machine Learning*, 2022.

[12] Kaggle. Give me some credit dataset. 2011.

[13] Karl Krauth, Yixin Wang, and Michael I. Jordan. Breaking feedback loops in recommender systems with causal inference, 2022.

[14] Qiang Li and Hoi-To Wai. State dependent performative prediction with stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3164–3186. PMLR, 2022.

[15] Daniel Liberzon. Calculus of variations and optimal control theory: A concise introduction. *Princeton University Press*, 2012.

[16] Chinmay Maheshwari, Chih-Yuan Chiu, Eric Mazumdar, Shankar Sastry, and Lillian 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.

[17] Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 2020.

[18] John Miller, Chloe Hsu, Jordan Troutman, Juan Perdomo, Tijana Zrnic, Lydia Liu, Yu Sun, Ludwig Schmidt, and Moritz Hardt. Whynot, 2020.

[19] John P. Miller, Juan C. Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. *CoRR*, abs/2102.08570, 2021.

[20] John P Miller, Juan C Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, pages 7710–7720. PMLR, 2021.

[21] Andjela Mladenovic, Iosif Sakos, Gauthier Gidel, and Georgios Piliouras. Generalized natural gradient flows in hidden convex-concave games and gans. In *International Conference on Learning Representations*, 2021.

[22] Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. *International Conference on Machine Learning*, 2020.

[23] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

[24] Mitas Ray, Lillian J Ratliff, Dmitriy Drusvyatskiy, and Maryam Fazel. Decision-dependent risk minimization in geometrically decaying dynamic environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[25] Alexey Tsymbal. The problem of concept drift: Definitions and related work. 05 2004.

6 Supplementary Material

6.1 Proof of Theorems

6.1.1 Auxiliary Lemmas

In order to prove theorems, we first state a definition and a lemma based on Chapter 1 of [15] 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 6.1 (*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

$$J(y + \alpha\eta) = J(y) + \delta J|_y(\eta)\alpha + o(\alpha). \tag{1}$$

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

$$\delta J|_y(\eta) = \lim_{\alpha\to 0} \frac{J(y + \alpha\eta) - J(y)}{\alpha}. \tag{2}$$

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 (2) 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:

$$g(\alpha) = g(0) + g'(0)\alpha + o(\alpha). \tag{3}$$

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 (3):

$$g(\alpha) - g(0) < g'(0)\alpha + |g'(0)\alpha|. \tag{4}$$

Since we assumed $g'(0) < 0$ and $\alpha > 0$, (4) 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 $\delta J|_y(\eta) \geq 0$. □

6.1.2 Proof of Theorem 2

For proving Theorem 2, we were inspired by the proof of Theorem 3.5 in [22], 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:

$$h(f_{\hat{\theta}}) = E_{z\sim\mathcal{D}(f_\theta)}[\ell(f_{\hat{\theta}}(x), y)] = \int \ell(f_{\hat{\theta}}(x), y)p_{f_\theta}(z)dz. \tag{5}$$

$$h'(f_{\hat{\theta}}) = E_{z\sim\mathcal{D}(f_{\theta'})}[\ell(f_{\hat{\theta}}(x), y)] = \int \ell(f_{\hat{\theta}}(x), y)p_{f_{\theta'}}(z)dz. \tag{6}$$

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:

$$\ell(f_{G(\theta)}(x), y) - \ell(f_{G(\theta')}(x), y) \geq \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right) \ell'(f_{G(\theta')}(x), y) + \frac{\gamma}{2} |f_{G(\theta)}(x) - f_{G(\theta')}(x)|^2. \tag{7}$$

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

$$h(f_{G(\theta)}) - h(f_{G(\theta')}) \geq \left(\int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right) \ell'(f_{G(\theta')}(x), y) p_{f_\theta}(z) dz \right) + \frac{\gamma}{2} \|f_{G(\theta)} - f_{G(\theta')}\|_\theta^2. \tag{8}$$

Similarly:

$$h(f_{G(\theta')}) - h(f_{G(\theta)}) \geq \left(\int \left(f_{G(\theta')}(x) - f_{G(\theta)}(x)\right) \ell'(f_{G(\theta)}(x), y) p_{f_\theta}(z) dz \right) + \frac{\gamma}{2} \|f_{G(\theta)} - f_{G(\theta')}\|_\theta^2. \tag{9}$$

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

$$\int \left(f_{G(\theta')}(x) - f_{G(\theta)}(x)\right) \ell'(f_{G(\theta)}(x), y) p_{f_\theta}(z) dz \geq 0. \tag{10}$$

to conclude:

$$-\gamma \|f_{G(\theta)} - f_{G(\theta')}\|_\theta^2 \geq \int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right) \ell'(f_{G(\theta')}(x), y) p_{f_\theta}(z) dz. \tag{11}$$

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

Now let's prove inequality (10) 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:

$$\begin{aligned}
\delta h|_{f_{G(\theta)}}(\eta) &= \lim_{\alpha \to 0} \frac{h(f_{G(\theta)} + \alpha\eta) - h(f_{G(\theta)})}{\alpha} \\
&= \lim_{\alpha \to 0} \int \frac{\ell(f_{G(\theta)}(x) + \alpha\eta(x), y) - \ell(f_{G(\theta)}(x), y)}{\alpha} p_{f_\theta}(z) dz \\
&= \int \lim_{\alpha \to 0} \frac{\ell(f_{G(\theta)}(x) + \alpha\eta(x), y) - \ell(f_{G(\theta)}(x), y)}{\alpha} p_{f_\theta}(z) dz \\
&= \int \lim_{\alpha \to 0} \frac{\ell(f_{G(\theta)}(x) + \alpha\eta(x), y) - \ell(f_{G(\theta)}(x), y)}{\alpha\eta(x)} \eta(x) p_{f_\theta}(z) dz \\
&= \int \ell'(f_{G(\theta)}(x), y) \eta(x) p_{f_\theta}(z) dz \\
&= \int \ell'(f_{G(\theta)}(x), y) (f_{G(\theta')}(x) - f_{G(\theta)}(x)) p_{f_\theta}(z) dz.
\end{aligned} \tag{12}$$

Knowing $\delta h|_{f_{G(\theta)}}(\eta) \geq 0$ completes the proof of (10).

Now recall that there exists M such that $M = \sup_{x,y,\theta} |\ell'(f_\theta(x), y)|$ and the distribution map over data is ϵ-sensitive w.r.t Pearson χ^2 divergence, i.e.

$$\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta)) \leq \epsilon \|f_\theta - f_{\theta'}\|^2. \tag{13}$$

With this in mind, we do the following calculations:

$$
\begin{aligned}
&\left| \int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right) \ell'(f_{G(\theta')}(x), y) p_{f_\theta}(z) dz - \int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right) \ell'(f_{G(\theta')}(x), y) p_{f_{\theta'}}(z) dz \right| \\
&= \left| \int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right) \ell'(f_{G(\theta')}(x), y) \left(p_{f_\theta}(z) - p_{f_{\theta'}}(z)\right) dz \right| \\
&\overset{(*)}{\leq} \int \left| \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right) \ell'(f_{G(\theta')}(x), y) \left(p_{f_\theta}(z) - p_{f_{\theta'}}(z)\right) \right| dz \\
&\leq M \int \left| \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right) \left(p_{f_\theta}(z) - p_{f_{\theta'}}(z)\right) \right| dz \\
&= M \int \left| \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right) \frac{p_{f_\theta}(z) - p_{f_{\theta'}}(z)}{p_{f_\theta}(z)} p_{f_\theta}(z) \right| dz \\
&= M \left| \int | \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right) \frac{p_{f_\theta}(z) - p_{f_{\theta'}}(z)}{p_{f_\theta}(z)} | p_{f_\theta}(z) dz \right| \\
&\overset{\text{Cauchy-Schwarz Ineq.}}{\leq} M \left(\int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right)^2 p_{f_\theta}(z) dz \right)^{\frac{1}{2}} \left(\int \left(\frac{p_{f_\theta}(z) - p_{f_{\theta'}}(z)}{p_{f_\theta}(z)} \right)^2 p_{f_\theta}(z) dz \right)^{\frac{1}{2}} \\
&= M \| f_{G(\theta)} - f_{G(\theta')} \|_\theta \sqrt{\chi^2(p_{f_\theta}, p_{f_{\theta'}})} \\
&\overset{13}{\leq} M \sqrt{\epsilon} \| f_{G(\theta)} - f_{G(\theta')} \|_\theta \| f_\theta - f_{\theta'} \|
\end{aligned}
$$

$(*)$ 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:

$$
\begin{aligned}
&\left| \int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right) \ell'(f_{G(\theta')}(x), y) p_{f_\theta}(z) dz - \int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right) \ell'(f_{G(\theta')}(x), y) p_{f_{\theta'}}(z) dz \right| \\
&\leq M \sqrt{\epsilon} \| f_{G(\theta)} - f_{G(\theta')} \|_\theta \| f_\theta - f_{\theta'} \|. \qquad (14)
\end{aligned}
$$

Similar to inequality (10), since $f_{G(\theta')}$ minimizes h', one can prove:

$$
\int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right) \ell'(f_{G(\theta')}(x), y) p_{f_{\theta'}}(z) dz \geq 0. \qquad (15)
$$

From (11) we know that $\int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right) \ell'(f_{G(\theta')}(x), y) p_{f_\theta}(z) dz$ is negative, so with this fact alongside (14) and (15), we can write:

$$
\int \left(f_{G(\theta)}(x) - f_{G(\theta')}(x)\right) \ell'(f_{G(\theta')}(x), y) p_{f_\theta}(z) dz \geq -M \sqrt{\epsilon} \| f_{G(\theta)} - f_{G(\theta')} \|_\theta \| f_\theta - f_{\theta'} \|. \qquad (16)
$$

Combining (11) and (16), we will get:

$$
\begin{aligned}
&\gamma \| f_{G(\theta)} - f_{G(\theta')} \|_\theta^2 \leq M \sqrt{\epsilon} \| f_{G(\theta)} - f_{G(\theta')} \|_\theta \| f_\theta - f_{\theta'} \| \\
&\Rightarrow \| f_{G(\theta)} - f_{G(\theta')} \|_\theta \leq \frac{\sqrt{\epsilon} M}{\gamma} \| f_\theta - f_{\theta'} \| \qquad (17)
\end{aligned}
$$

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

$$
\| f_{G(\theta)} - f_{G(\theta')} \|^2 \leq C \| f_{G(\theta)} - f_{G(\theta')} \|_\theta^2 \qquad (18)
$$

Consequently,

$$
\| f_{G(\theta)} - f_{G(\theta')} \| \leq \sqrt{C} \| f_{G(\theta)} - f_{G(\theta')} \|_\theta \qquad (19)
$$

Using (19) in (17) results in:

$$
\| f_{G(\theta)} - f_{G(\theta')} \| \leq \frac{\sqrt{C\epsilon} M}{\gamma} \| f_\theta - f_{\theta'} \|. \qquad (20)
$$

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:

$$\begin{aligned}\|f_{\theta_t} - f_{\theta_{PS}}\| &\leq \frac{\sqrt{C\epsilon}M}{\gamma}\|f_{\theta_{t-1}} - f_{\theta_{PS}}\| \\ &\leq (\frac{\sqrt{C\epsilon}M}{\gamma})^t\|f_{\theta_0} - f_{\theta_{PS}}\| \end{aligned} \tag{21}$$

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 (21),

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

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

6.1.3 Proof of Theorem 3

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

$$\begin{aligned} p_{f_\theta}(x) &= p(x)\Big(1 - g(f_\theta(x))\Big) + p(x)\mathbb{E}_X[g(f_\theta(X))] \\ &= p(x)(1 - g(f_\theta(x)) + C_\theta). \end{aligned} \tag{22}$$

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:

$$p_{f_\theta}(x) = p(x)(1 - f_\theta(x) - \delta + C_\theta). \tag{23}$$

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. However, we assume that the label is a deterministic function of the features, so for $(x, y) \sim p(z)$, $p(y|x) = 1$ for the true label y, and this simplifies our following calculations since $p_{f_\theta}(z) = p_{f_\theta}(x)$ and $p(z) = p(x)$ for data points that have a positive probability.

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

$$\begin{aligned}
\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta)) &= \int \frac{(p_{f_{\theta'}}(x) - p_{f_\theta}(x))^2}{p_{f_\theta}(x)}dx \\
&= \int \frac{p(x)^2(f_\theta(x) - f_{\theta'}(x) - (C_\theta - C_{\theta'}))^2}{p(x)(1 - f_\theta(x) - \delta + C_\theta)}dx \\
&\overset{C_\theta \geq \delta}{\leq} \frac{1}{\delta}\int p(x)\Big[(f_\theta(x) - f_{\theta'}(x))^2 + (C_\theta - C_{\theta'})^2 - 2(f_\theta(x) - f_{\theta'}(x))(C_\theta - C_{\theta'})\Big]dx \\
&= \frac{1}{\delta}\Big[\Big(\int p(x)(f_\theta(x) - f_{\theta'}(x))^2dx\Big) + (C_\theta - C_{\theta'})^2 - 2(C_\theta - C_{\theta'})\int p(x)(f_\theta(x) - f_{\theta'}(x))dx\Big] \\
&\overset{(*')}{=} \frac{1}{\delta}\Big[\Big(\int p(x)(f_\theta(x) - f_{\theta'}(x))^2dx\Big) + (C_\theta - C_{\theta'})^2 - 2(C_\theta - C_{\theta'})^2\Big] \\
&= \frac{1}{\delta}\Big[\int p(x)(f_\theta(x) - f_{\theta'}(x))^2dx - (C_\theta - C_{\theta'})^2\Big] \\
&\leq \frac{1}{\delta}\int p(x)(f_\theta(x) - f_{\theta'}(x))^2dx \\
&= \frac{1}{\delta}\|f_\theta - f_{\theta'}\|^2
\end{aligned} \tag{24}$$

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:

$$\frac{p(x)}{p_{f_{\theta^*}}(x)} = \frac{1}{1 - g(f_{\theta^*}(x)) + C_{\theta^*}} \leq \frac{1}{\delta}. \tag{25}$$

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 norm 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:

$$p_{f_\theta}(x) = p(x)(1 - g(f_\theta(x))) + \int_{x'} p(x'_s, x'_f = x_f) \; g(f_\theta(x')) \; p(x_s|x_f)dx' \tag{26}$$

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 (26) as follows:

$$\begin{aligned}
p_{f_\theta}(x) =& p(x)(1 - g(f_\theta(x))) + \int_{x'} p(x'_s, x'_f = x_f) \; g(f_\theta(x')) \; p(x_s|x_f)dx' \\
&= p(x)(1 - g(f_\theta(x))) + \int_{x'} g(f_\theta(x'))p_{X_s}(x'_s)p_{X_f}(x_f)p_{X_s}(x_s)dx' \\
&= p(x)(1 - g(f_\theta(x))) + \int_{x'} g(f_\theta(x'))p_{X_s}(x'_s)p(x)dx' \\
&= p(x)\Big((1 - g(f_\theta(x))) + \int_{x'} g(f_\theta(x'))p_{X_s}(x'_s)dx'\Big)
\end{aligned} \tag{27}$$

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 (22) 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. □

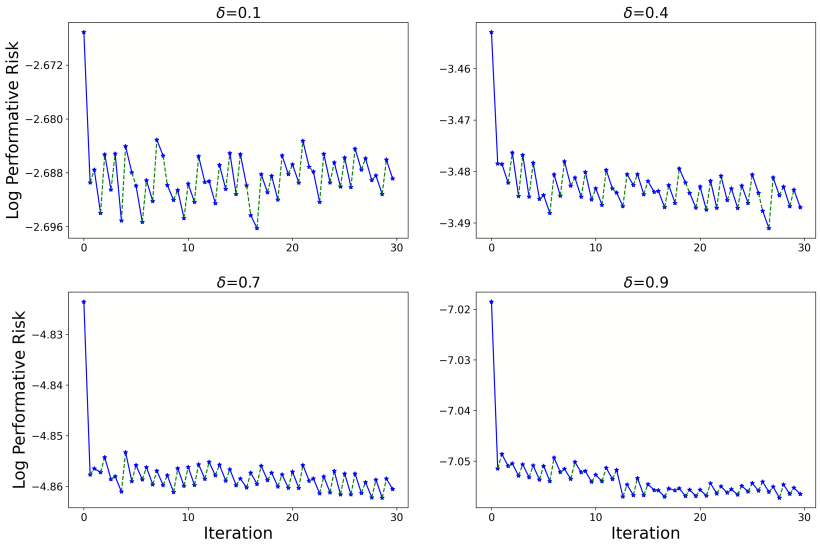

Figure 2: Evolution of log of performative risk for different values of $\delta = 0.1, 0.4, 0.7, 0.9$ through iterations of RRM.

6.2 Additional Experiments

Figures 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 our theory, 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.