Performative Prediction with Neural Networks

Anonymous Author(s)

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 [8] to deal with the problem of distribution shift or concept drift ([2, 10, 9]) 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, traffic forecasts, and recommender systems ([8, 4]).

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 ([8, 6, 1]). 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 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. We believe this is a significant step toward the generalization of the performative prediction framework, as we can now theoretically analyze the performative effects of a subset of neural networks with non-convex loss functions.

Submitted to the Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

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 [8] 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 $\mathcal{D}(f_\theta)$. Following is the informal statement of our main theorem.

Theorem 1. *(Informal) If the loss $\ell(f_\theta(z), z)$ is strongly convex in $f_\theta(z)$ 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 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 [8] 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 ([8, 7, 1]). 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

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 $\theta \in \Theta$. In this new formulation, the objective risk function which we call *Functional Performative Risk (FPR)* is defined as

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

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:

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

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

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

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*) A distribution map $\mathcal{D}(.)$ which maps f_θ to $\mathcal{D}(f_\theta)$ with probability density function (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:

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

where $\|f_\theta - f_{\theta'}\|_\theta^2 = \int |f_\theta(z) - f_{\theta'}(z)|^2 p_{f_\theta} dz$ 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 ratio*) A distribution map $\mathcal{D}(.)$ satisfies bounded ratio with the parameter $C > 1$ if for all $f_\theta, f_{\theta'} \in \mathcal{F}$:

$$\forall z, p_{f_\theta}(z) \leq C\, p_{f_{\theta'}}(z)$$

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

A3 (*Strong convexity w.r.t predictions*) A loss function $\ell(f_\theta(z), z)$ which takes as inputs an instance z and its prediction under function f_θ, is γ-strongly convex in $f_\theta(z)$ if the following inequality holds for every $\theta, \theta' \in \Theta$:

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

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

We can easily see that these two assumptions on ℓ are satisfied by the Squared Error loss: $\ell(f_\theta(z), z) = \frac{1}{2}(f_\theta(x) - y)^2$, which is 1-strongly convex with a bounded derivative of 1 if $y \in \{0, 1\}$ and $f_\theta(x) \in [0, 1]$ for any θ.

Convergence of FRRM Now we can state our main theorem which provides sufficient conditions for functional repeated risk minimization to converge to a stable classifier with unique predictions.

Theorem 2. *Suppose that the loss $\ell(f_\theta(z), z)$ is γ-strongly convex w.r.t $f_\theta(z)$ (A3) and its derivative w.r.t $f_\theta(z)$ is bounded with $M = \sup_{z,\theta} |\ell'(f_\theta(z))|$ (A4). If the distribution map $\mathcal{D}(.)$ is ϵ-sensitive w.r.t Pearson χ^2 divergence (A1) and satisfies Bounded ratio with parameter C(A2), then:*

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

So if $\frac{\sqrt{C\epsilon}M}{\gamma} < 1$, G is a contractive mapping and FRRM converges to a stable point at a linear rate: $\|f_{\theta_t} - f_{\theta_{PS}}\|_{\theta_t} \leq \alpha$ for $t \geq (1 - \frac{\sqrt{C\epsilon}M}{\gamma})^{-1} \log(\frac{\|f_{\theta_0} - f_{\theta_{PS}}\|_{\theta_0}}{\alpha})$

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

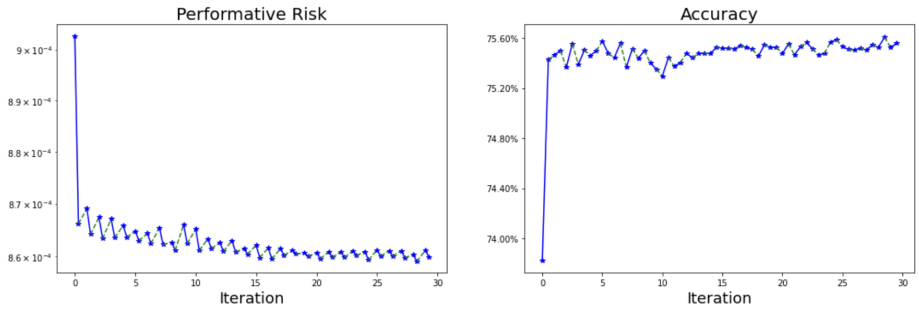

Figure 1: Evolution of performative risk (left) and accuracy (right) through iterations of FRRM for $\delta = 0.9$. The blue lines show the changes in risk (accuracy) after optimizing on the distribution induced by the last model, and the green lines show the effect of the distribution shift on the risk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

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 RIR(f_θ) 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 take a 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 $p_{f_\theta}(x) = p(x)\Big(1 - g(f_\theta(x))\Big) + p(x) \int p(y) g(f_\theta(y)) dy$

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 $\frac{1}{\delta^2}$-sensitive w.r.t χ^2 divergence (A1) and satisfies the Bounded ratio assumption (A2) for $C = \frac{2-\delta}{\delta}$.*

4 Experiments

We complement our theoretical results with experiments on a credit-scoring task and illustrate how they support our claims. 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 ([3]), 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.

In our simulations, we assume that the data distribution induced by the classifier f_θ shifts according to the RIR procedure where strategic features—those that can be (easily) manipulated without affecting the true label, e.g. Number of open credit lines and loans—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. Remember that $y = 1$ corresponds to default, so 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. $\mathbb{E}[\frac{1}{2}(f_\theta(x) - y)^2]$.

Figure 1 shows the evolution of performative risk (left) and accuracy (right) through iterations of FRRM 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, FRRM converges to a stable model; this is supported by our results in Figure 1.

References

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

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

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

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

[5] Daniel Liberzon. Calculus of variations and optimal control theory: A concise introduction. 2012.

[6] Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33:4929–4939, 2020.

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

[8] Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. 2021.

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

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

5 Supplementary Material

5.1 Proof of Theorem 2

Proof. Fix θ and θ' in Θ. Let h and h' be two functionals defined as follows:

$$h(f_{\hat{\theta}}) = E_{z\sim p_{f_\theta}}[\ell(f_{\hat{\theta}}(z))] = \int \ell(f_{\hat{\theta}}(z))p_{f_\theta}(z)dz \tag{1}$$

$$h'(f_{\hat{\theta}}) = E_{z\sim p_{f_{\theta'}}}[\ell(f_{\hat{\theta}}(z))] = \int \ell(f_{\hat{\theta}}(z))p_{f_{\theta'}}(z)dz \tag{2}$$

For a fixed z, due to strong convexity of ℓ w.r.t $f(z)$ we have:

$$\ell(f_{G(\theta)}(z)) - \ell(f_{G(\theta')}(z)) \geq \left(f_{G(\theta)}(z) - f_{G(\theta')}(z)\right)\ell'(f_{G(\theta')}(z)) + \frac{\gamma}{2}|f_{G(\theta)}(z) - f_{G(\theta')}(z)|^2 \tag{3}$$

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

$$h(f_{G(\theta)}) - h(f_{G(\theta')}) \geq \left(\int \left(f_{G(\theta)}(z) - f_{G(\theta')}(z)\right)\ell'(f_{G(\theta')}(z))p_{f_\theta}(z)dz\right) + \frac{\gamma}{2}\|f_{G(\theta)} - f_{G(\theta')}\|_\theta^2 \tag{4}$$

Similarly:

$$h(f_{G(\theta')}) - h(f_{G(\theta)}) \geq \left(\int \left(f_{G(\theta')}(z) - f_{G(\theta)}(z)\right)\ell'(f_{G(\theta)}(z))p_{f_\theta}(z)dz\right) + \frac{\gamma}{2}\|f_{G(\theta)} - f_{G(\theta')}\|_\theta^2 \tag{5}$$

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

$$\int \left(f_{G(\theta')}(z) - f_{G(\theta)}(z)\right)\ell'(f_{G(\theta)}(z))p_{f_\theta}(z)dz \geq 0 \tag{6}$$

to conclude:

$$-\gamma\|f_{G(\theta)} - f_{G(\theta')}\|_\theta^2 \geq \int \left(f_{G(\theta)}(z) - f_{G(\theta')}(z)\right)\ell'(f_{G(\theta')}(z))p_{f_\theta}(z)dz \tag{7}$$

Before proving 6, we need a definition and a lemma which are proposed below, based on [5].

Definition 5.1 (*First variation of a functional*)**.** *Let $J : V \to R$ be a functional on a function space V, and consider some function $y \in V$. The derivative of J at y is called the first variation (a.k.a Gateaux derivative), and is defined as below:*
A linear functional $\delta J|_y : V \to R$ is called the first variation of J at y if for all η and all α we have

$$J(y + \alpha\eta) = J(y) + \delta J|_y(\eta)\alpha + o(\alpha) \tag{8}$$

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

$$\delta J|_y(\eta) = \lim_{\alpha\to 0}\frac{J(y + \alpha\eta) - J(y)}{\alpha} \tag{9}$$

Lemma 1 (*First-order necessary condition for optimality in a constrained function space*)**.** *If y is a minimum point 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 9 we can conclude $\delta J|_y(\eta) = g'(0)$, so it suffices to show that $g'(0) \geq 0$.
Since y is a minimum point of J, then 0 is a minimum point of g. The first-order Taylor approximation of g around 0 is as follows:

$$g(\alpha) = g(0) + g'(0)\alpha + o(\alpha) \tag{10}$$

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 ineuqality using 10:

$$g(\alpha) - g(0) < g'(0)\alpha + |g'(0)\alpha| \tag{11}$$

Since we assumed $g'(0) < 0$ and $\alpha > 0$, 11 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 $\delta J|_y(\eta) \geq 0$. □

Now let's prove inequality 6. 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 minimum point 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:

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

Knowing $\delta h|_{f_{G(\theta)}}(\eta) \geq 0$ completes the proof of 6.

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

$$\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta)) = \int \frac{\left(p_{f_{\theta'}}(z) - p_{f_\theta}(z)\right)^2}{p_{f_\theta}(z)} dz \leq \epsilon \|f_\theta - f_{\theta'}\|_\theta^2 \tag{13}$$

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

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

What we can conclude from the above derivations is that:

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

Similar to inequality 6, since $f_{G_{\theta'}}$ minimizes h', one can prove:

$$\int \left(f_{G(\theta)}(z) - f_{G(\theta')}(z)\right) \ell'(f_{G(\theta')}(z)) p_{f_{\theta'}}(z)dz \geq 0 \tag{15}$$

From 7 we know that $\int \left(f_{G(\theta)}(z) - f_{G(\theta')}(z)\right) \ell'(f_{G(\theta')}(z)) p_{f_\theta}(z) dz$ is negative, so with this fact and 15, we will have:

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

Combining 7 and 16, we'll get:

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

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

$$\begin{aligned} \|f_{G(\theta)} - f_{G(\theta')}\|_{G(\theta)}^2 &= \int |f_{G(\theta)}(z) - f_{G(\theta')}(z)|^2 p_{f_{G(\theta)}}(z) dz \\ &\leq C \int |f_{G(\theta)}(z) - f_{G(\theta')}(z)|^2 p_{f_\theta}(z) dz \\ &= C\|f_{G(\theta)} - f_{G(\theta')}\|_\theta^2 \end{aligned} \tag{18}$$

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

Using 19 in 17 results in:

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

So if $\frac{\sqrt{C\epsilon}M}{\gamma} < 1$, G is a contractive mapping and FRRM would converge to a stable point.

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

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

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

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

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

□

5.2 Proof of Theorem 3

Proof. The pdf of $p_{f_\theta}(x)$ is as follows:

$$\begin{aligned} p_{f_\theta}(x) &= p(x)\left(1 - g(f_\theta(x))\right) + p(x)\int p(y)g(f_\theta(y))dy \\ &= p(x)(1 - g(f_\theta(x)) + C_\theta) \end{aligned} \tag{22}$$

where $C_\theta = \int p(y)g(f_\theta(y))dy$.
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 $\delta \leq g(f_\theta(x)) \leq 1$ for every x.

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

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

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

$$\begin{aligned}
\int p(x)(f_\theta(x) - f_{\theta'}(x))^2 dx &= \int \frac{p(x)(1 - f_\theta(x) - \delta + C_\theta)(f_\theta(x) - f_{\theta'}(x))^2}{1 - f_\theta(x) - \delta + C_\theta} dx \\
&= \int \frac{p_{f_\theta}(x)(f_\theta(x) - f_{\theta'}(x))^2}{1 - f_\theta(x) - \delta + C_\theta} dx \\
&\leq \frac{1}{\delta} \int p_{f_\theta}(x)(f_\theta(x) - f_{\theta'}(x))^2 dx \\
&= \frac{1}{\delta} \|f_\theta - f_{\theta'}\|_\theta^2
\end{aligned} \tag{25}$$

Integrating 24 and 25 Will result in ϵ-sensitivity with $\epsilon = \frac{1}{\delta^2}$:

$$\chi^2(\mathcal{D}(f_{\theta'}), \mathcal{D}(f_\theta)) \leq \frac{1}{\delta^2} \|f_\theta - f_{\theta'}\|_\theta^2 \tag{26}$$

distribution map is ϵ-sensitive w.r.t χ^2 divergence for $\epsilon = \frac{1}{\delta^2}$.

Since $\delta \leq C_\theta \leq 1$, it is easy to see that for any f_θ and $f_{\theta'}$ and for any x:

$$p_{f_\theta}(x) \leq \frac{2 - \delta}{\delta} p_{f_{\theta'}}(x) \tag{27}$$

So the distribution map satisfies the bounded ratio condition for $C = \frac{2-\delta}{\delta}$. $\square$