Which Echo Chamber?
Regions of Attraction in Learning with Decision-Dependent Distributions

Roy Dong
Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Champaign, IL
roydong@illinois.edu

Lillian J. Ratliff
Electrical and Computer Engineering
University of Washington
Seattle, WA
ratliffl@uw.edu

Abstract

As data-driven methods are deployed in real-world settings, the processes that generate the observed data will often react to the decisions of the learner. For example, a data source may have some incentive for the algorithm to provide a particular label (e.g. approve a bank loan), and manipulate their features accordingly. Work in strategic classification and decision-dependent distributions seeks to characterize the closed-loop behavior of deploying learning algorithms by explicitly considering the effect of the classifier on the underlying data distribution. More recently, works in performative prediction seek to classify the closed-loop behavior by considering general properties of the mapping from classifier to data distribution, rather than an explicit form. Building on this notion, we analyze repeated risk minimization as the perturbed trajectories of the gradient flows of performative risk minimization. We consider the case where there may be multiple local minimizers of performative risk, motivated by real world situations where the initial conditions may have significant impact on the long-term behavior of the system. As a motivating example, we consider a company whose current employee demographics affect the applicant pool they interview: the initial demographics of the company can affect the long-term hiring policies of the company. We provide sufficient conditions to characterize the region of attraction for the various equilibria in this settings. Additionally, we introduce the notion of performative alignment, which provides a geometric condition on the convergence of repeated risk minimization to performative risk minimizers.

1 Introduction

Data-driven methods are growing increasingly popular in practice. Most classical machine learning and statistical methods view the underlying process which generates the data as fixed: the study is primarily focused on the mapping from data distributions to classifier. However, it is important to consider the effects in the other direction as well: how does the classifier chosen by a learner change the data distribution the learner sees? In particular, how do we close the loop around machine learning deployments in practice?

These closed loop effects can arise in many real world settings. One instance is strategic classification: whenever a data source has a stake in which label a classifier applies to it, they will seek cost-effective ways to manipulate their data to earn the desired label. For example, credit scoring classifiers are heavily guarded for fear of the potential for gaming [Hardt et al., 2016]. Alternatively, deployments of the classifier can both skew future datasets and also have causal influences over the real-world

Preprint. Under review.

processes at play. For example, a classifier that predicts crime recidivism influences the opportunities available to individuals [Dressel and Farid, 2018].

Formally, we consider this problem in the framework introduced in Perdomo et al. [2020]. Let $\ell(z, x)$ denote the loss when the learner's decision is x (e.g. x can be the parameters of the chosen classifier) and the data has realized value z. Furthermore, let $\mathcal{D}(x)$ denote the data distribution when the learner's decision is x. In this framework, the performative risk is given by:

$$PR(x) = \mathbb{E}_{Z \sim \mathcal{D}(x)}[\ell(Z, x)] \tag{1}$$

Whereas classical machine learning results treat the distribution $Z \sim \mathcal{D}$ as fixed, the performative prediction framework models the decision-dependent distribution as a mapping $\mathcal{D}(\cdot)$. However, in many real world-deployments, this decision-dependent distribution shift may not be explicitly included in the learner's updates. This leads to algorithms based on inexact repeated minimization. Define the decoupled performative risk as:

$$R(x_1, x_2) = \mathbb{E}_{Z \sim \mathcal{D}(x_2)}[\ell(Z, x_1)] \tag{2}$$

The decoupled performative risk $R(x_1, x_2)$ separates the two ways that the decision variable x affects the performative risk. Through the x_1 argument, x affects the classification error; through the x_2 argument, x causes a decision-dependent distribution shift. In this paper, we shall analyze the steady-state behavior of stochastic gradient descent algorithms:

$$x_{k+1} = x_k - \alpha_k(\nabla_{x_1} R(x_k, x_k) + \eta_k) \tag{3}$$

Here, $(\eta_k)_k$ is some zero-mean noise process. Note that the gradient is evaluated only with respect to the first argument, i.e. the updates are based only on the effect of x on the loss function, and ignore the distribution shift caused by x. In other words, the learner draws several observations from the distribution $\mathcal{D}(x_k)$, and, treating this distribution as fixed, updates their model parameters x_{k+1} based on stochastic gradient descent: they are descending the gradient of the cost function $y \mapsto R(y, x_k)$.

In particular, we focus on settings where there may be multiple local equilibria, and classify their regions of attraction for these equilibria. In many settings of interest, there may be multiple steady-state outcomes, and it is of interest to determine which outcome will be chosen by the dynamics in Equation (3). As a motivating example, we consider a model of how a company's demographics can affect the pool of applications that apply for jobs at the company. In this model, the initial demographics of the company determine the steady-state demographics of the company. Our results allow us to characterize which regions of the parameter space will converge to which equilibria. We discuss this example in greater formal detail in Section 3.1.

Our main theoretical results can be informally summarized as follows. Theorem 1 states that trajectories of inexact repeated risk minimization will converge exponentially fast to a neighborhood of local performative risk minimizers, and stay in this neighborhood for all future time. It also provides a sufficient condition to under-approximate the regions of attraction for each local performative risk minimizer. In the special case of vanishing perturbations, these trajectories will converge to the minimizers themselves. As a corollary, this implies that performatively stable points will be near performatively optimal points, which can be seen as a continuous-time analog to results proved in Perdomo et al. [2020]. Theorem 2 states a geometric condition on the performative perturbation which ensures that trajectories of repeated risk minimization will converge to local performative risk minimizers, intuitively based on the idea that the perturbation does not push against convergence.

These results allow us to identify the regions of attraction for various steady-state outcomes. As observed in Miller et al. [2021], these various outcomes can be interpreted as different echo chambers: essentially the decision variable x can act as a sort of self-fulfilling prophecy.[1] In settings with multiple echo chambers, we consider the question of which echo chamber will come to dominate, based on the initialization of the learner.

The rest of the paper is organized as follows. In Section 2, we discuss the related literature. In Section 3, we introduce the problem statement and the mathematical concepts used for our results, and

[1] It is worth noting that we take a slightly different interpretation of an 'echo chamber' in this paper. In Miller et al. [2021], the echo chambers are defined as performatively stable points. In this paper, we consider the regions near each locally performatively optimal point as an echo chamber. As we will discuss in Section 3.1, we are interested in settings where there may be many local performative risk minimizers that attract learning methods depending on initialization.

provide a motivating example based on job applicant pools in Section 3.1. In section 4, we analyze the gradient flow associated with performative risk minimization, and in Section 5, we analyze the flows associated with repeated risk minimization. We demonstrate numerical results in Section 6, and provide closing remarks in Section 7.

2 Background

There has been a great deal of interest in studying decision-dependent distributions. In the context of operations research, this has been studied under either the name decision-dependent uncertainty or endogenous uncertainty. In Jonsbråten et al. [1998], Jonsbråten [1998], and Goel and Grossmann [2004], the authors considered oil field optimization, with a framework that captures how information revelation can be affected by one's decisions. In Peeta et al. [2010], the authors consider infrastructure investment, and how investments can affect the future likelihood of disasters. For a taxonomy of the work in the operations research community, we refer the reader to Hellemo et al. [2018].

Another form of decision-dependent distributions is strategic classification. In these works, the data source is seen as a utility-maximizing agent. The distribution shift resulting from the learner's decision is modeled by a best response function. In Hardt et al. [2016] and Brückner and Scheffer [2011], the authors formulate the problem as a Stackelberg game where the data source responds to the announced classifier. In Dong et al. [2018], the authors consider when the data source's preferences are hidden information and provide sufficient conditions for convexity of the overall strategic classification task. In Akyol et al. [2016], the authors quantify the cost of strategic classification for the classifier. In Milli et al. [2019] and Hu et al. [2019], the authors note that certain groups may be disproportionately affected as institutions incorporate methods to counter data sources gaming the classifier. In Miller et al. [2020], the authors formulate strategic classification in a causal framework.

Most related to our work is recent efforts in performative prediction. This was introduced in [Perdomo et al., 2020]. In this formulation, rather than explicitly modeling the form of the distribution shift, it proposes to analyze the decision-dependent distribution shift in terms of general properties of the $\mathcal{D}(\cdot)$ mapping, where $\mathcal{D}(x)$ is the distribution of the data when the learner's decision is x. In Perdomo et al. [2020], the authors introduced the concepts related to performative prediction, demonstrated that neither the performatively stable nor performatively optimal points are subsets of each other, provided sufficient conditions for exact repeated risk minimization (defined as finding the exact minima with respect to $\mathcal{D}(x_k)$ at each time step) to converge, and provided conditions in which performatively stable points are near performatively optimal points. In Mendler-Dünner et al. [2020], the authors analyze inexact repeated risk minimization (defined as an update step with respect to $\mathcal{D}(x_k)$ at each time step) from a stochastic optimization framework. In this paper, we build on the inexact repeated risk minimization framework. Miller et al. [2021] provided sufficient conditions for performative risk itself to be convex. Brown et al. [2020] extended these results to settings where the distribution updates may have an internal state. In Drusvyatskiy and Xiao [2020], the authors show that many inexact repeated risk minimization algorithms will also converge nicely, due to the way in which the performative perturbation decays near the solution. This shares many ideas with our work here, but we focus on the case where there may be multiple attractive equilibria, and generalize to settings where the perturbation itself may not vanish. In contrast to previous works which provide sufficient conditions to guarantee that an outcome is approached globally, we focus on understanding local regions of attraction for various outcomes.

This work draws on ideas from control theory; in particular, the analysis of gradient flows, Lyapunov functions, and perturbation analysis are the tools we use throughout. We refer the reader to Hirsch et al. [2012] and Khalil [2001] as good references for these suite of tools.

3 Performative prediction, flows, and perturbations

In this section, we introduce the mathematical concepts used throughout this paper. As previously mentioned, the framework used throughout this paper builds on the framework of performative prediction, introduced in Perdomo et al. [2020].

In Section 1, we have already defined the **performative risk** in Equation (1) and the **decoupled performative risk** in Equation (2). Furthermore, we say that x is a **local performative risk minimizer** is x is a local minima of $PR(\cdot)$. We say x is **locally performatively stable** if x is a local minima of $y \mapsto R(y, x)$. In general, neither imply the other [Perdomo et al., 2020].

Additionally, we consider the **performative risk minimizing (PRM) gradient flow**, defined by the following differential equation:

$$\dot{x}_{PR} = -\nabla PR(x_{PR}) = -\nabla_{x_1} R(x_{PR}, x_{PR}) - \nabla_{x_2} R(x_{PR}, x_{PR}) =: f_{PR}(x_{PR}) \tag{4}$$

This vector field can be represented by the gradient of a function, which lends the flow to nice analysis. Under mild conditions, the trajectories of Equation (4) will converge to local minima of the performative risk.

However, as noted in Section 1, many deployments of machine learning do not explicitly model the distribution shift, and, consequently, do not directly minimize the performative risk. We define the **repeated risk minimizing (RRM) flow** as solutions to the differential equation:

$$\dot{x}_{RR} = -\nabla_{x_1} R(x_{RR}, x_{RR}) =: f_{RR}(x_{RR}) \tag{5}$$

We define the **performative perturbation**:

$$g(x) := \nabla_{x_2} R(x, x) = f_{RR}(x) - f_{PR}(x)$$

In this paper, we view the PRM gradient flow as the *nominal* dynamics, and the RRM flow as the *perturbed* dynamics. The PRM gradient flow has nice properties arising from the fact it is a gradient flow, and, under certain conditions on the performative perturbation, we can prove properties about the RRM flow, which is the quantity of interest. In particular, we show ultimate bounds on the distance between the trajectories of RRM flow and the local performative risk minimizers. This also implies that under certain conditions on the performative risk, all performatively stable points are near performative risk minimizers, as was observed in Perdomo et al. [2020].

Throughout this paper, we will be using tools from perturbation analysis in control theory. For a complete vector field $\dot{x} = f(x)$, let $\varphi_f(\cdot; x_0)$ denote the unique solution to the differential equation with initial condition $x(0) = x_0$. For a scalar-valued function V and a vector field f, we can define the derivative along trajectories as $\mathcal{L}_f V(x) = \frac{\partial V}{\partial x} f(x)$. We say a point x is an equilibrium point if $f(x) = 0$. An equilibrium point x is **locally asymptotically stable** if there exists a neighborhood $U \ni x$ such that $\lim_{t\to\infty} \varphi_f(t; x') = x$ for all $x' \in U$. A set A is **positively invariant** if for all $x_0 \in A$ and $t \geq 0$, we have $\varphi_f(t; x_0) \in A$. Additionally, given a set $A \subset \mathbb{R}^n$, we say two points x and y are **path-connected in** A if there exists a continuous function $\gamma : [0, 1] \to A$ such that $\gamma(0) = x$ and $\gamma(1) = y$. This forms an equivalence relation defined on A, and each equivalence class is a **connected component of** A.

3.1 A motivating example: decision-dependent distribution shift in job applicant pools

Before we present our analysis of the PRM gradient flow and the RRM flow, we introduce a motivating example which motivates the study of multiple local equilibria. This toy example considers a model for decision-dependent distribution shift in the applicant pool for a job based on the past hiring decisions of a company.

In this model, individuals are characterized by three variables: 1) their group membership, $\theta \in \{0, 1\}$, 2) $y \sim N(0, 1)$, which is their true productive capacity and independent of their group membership θ, and 3) $\hat{y} \sim N(y, \sigma^2)$, their observable productivity, which is a distorted version of their true productive capacity. The distribution of θ in the applicant pool will be decision-dependent, which we shall specify shortly. We let $z = (\theta, \hat{y}, y)$, which is viewed as observable features $(\theta, \hat{y})$ and output y.

A learner wants to use historical hiring data to determine which individuals to hire into their company. The variable of interest is y, which is the individual's true productive capacity. The learner must make the decision based on $(\theta, \hat{y})$, and can observe y after the fact. We assume the learner uses a linear classifier. Let $f(\theta, \hat{y}; x) = x_b + x_\theta \theta + x_{\hat{y}} \hat{y}$, and the individual with observable features $(\theta, \hat{y})$ is hired if $f(\theta, \hat{y}; x) \geq 0$. We also assume the learner uses a logistic loss function:

$$\ell(z, x) = \log(1 + \exp(-y f(\theta, \hat{y}; x)))$$

Now, let us define the decision-dependent distribution shift $\mathcal{D}(x)$. In our model, if group $\theta = 1$ has insufficient representation among the hired population, then future rounds will have fewer applications

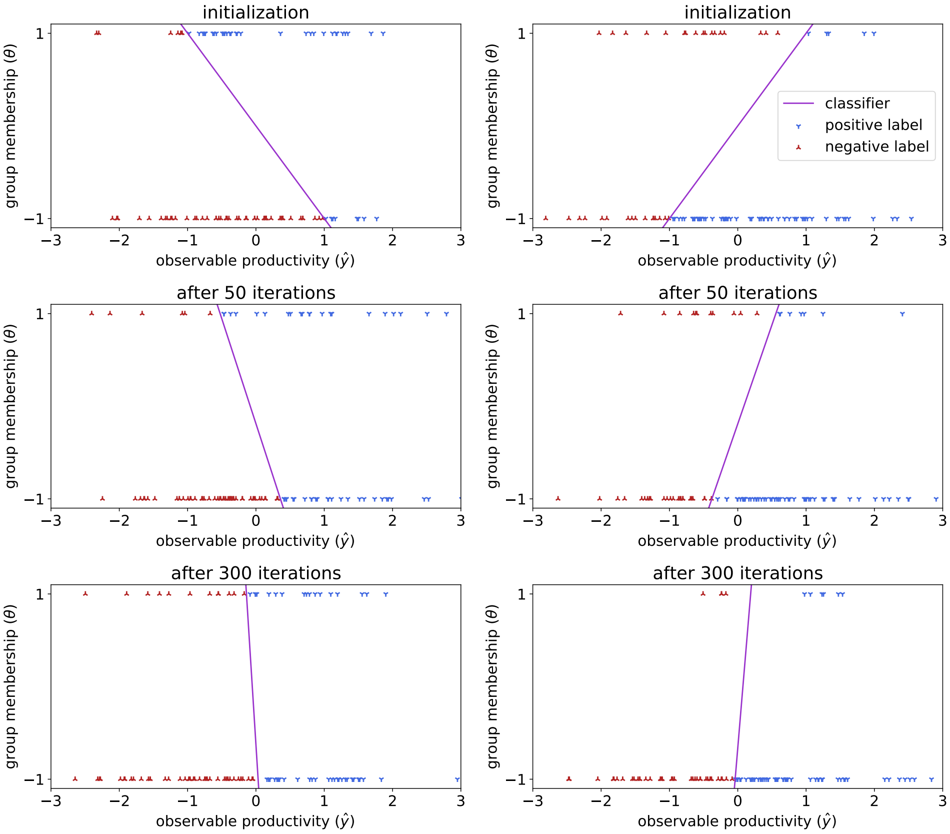

Figure 1: The population and classifier across time. On the left, we see an initialization that favors group 1. After 300 iterations, we see that the classifier converges to primarily only using the $\hat{y}$ feature, and that group 1 is still represented in the applicant pool. On the right, we see an initialization that is unfavorable to group 1. After 300 iterations, the classifier still converges and ignores the group membership, but the participation of group 1 in the applicant pool has dwindled.

from this group. Furthermore, if group $\theta = 1$ is well-represented in the hired population, then this encourages this group to apply in the future. We suppose there is some critical fraction p_c: if less than p_c of the hired population is of group 1, the next applicant pool will have reduced participation by group 1; if the hired population has group 1 represented by more than p_c, then the next applicant pool will have increased participation up to a saturation fraction $\overline{p}$.

Formally, this means we define $\mathcal{D}(x)$ as follows. The distributions of y and $\hat{y}$ do not depend on x. Let p denote the fraction of the applicant pool that was group 1 in the previous iteration, and let q denote the fraction of the previously accepted applicants which was group 1. The group membership $\theta \sim Bernoulli(p^+)$, where p^+ is given by:[2]

$$p^+ = \text{median}(0, \overline{p}, p - \beta q(q - p_c)(q - \overline{p})) \tag{6}$$

This update rule decreases p for $q \in (0, p_c)$ and increases p for $q \in (p_c, \overline{p})$.

We show numerical some of the results in Figure 1. In this example, we set the critical fraction as $p_c = 0.3$, the saturation fraction $\overline{p} = 0.5$, and the observation variance $\sigma^2 = (0.8)^2$. We initialized $p = p_c = 0.3$, and, at each iteration k, the dataset was redrawn with 100 samples from $\mathcal{D}(x_k)$. In this model, since θ is uncorrelated with the quantity of interest y, we see the classifier converges to a vertical line which ignores the group membership θ in both cases. However, in the initialization that

[2] We note that, technically, this formulation requires the $\mathcal{D}$ mapping have some notion of 'state', since the next distribution depends on not only x but the fraction of the previous applicant pool p and accepted applicants q. Such extensions have been considered generally in Brown et al. [2020], but, for this example, it suffices to add p and q as components of x and z, since both quantities are known by the learner.

is unfavorable to group 1, we see a dwindling participation by group 1, even though the classifier after 300 iterations is relatively fair.[3]

This model motivates the study of regions of attraction for different equilibria. Our results can identify the region of attraction for different outcomes. A company for one reason or another may have historically hired more from one group of individuals than another. This example shows that when there are decision-dependent distribution shifts, the initial conditions can affect the final outcomes. In particular, for this setting, it is of interest to identify the region of convergence for different equilibria.

4 Analysis of performative risk minimizing gradient flow

In this section, we consider PRM gradient flow, defined by Equation (4). We observe that gradient flows provide complete vector fields, and that trajectories will converge to local performative risk minimizers under very mild conditions.

First, we state a proposition guaranteeing that flow is well-defined. The compact sublevel sets ensure that trajectories of Equation (4) remain bounded, which is sufficient to guarantee existence and uniqueness of solutions globally. For proof of the following proposition, we refer the reader to either Khalil [2001, Section 3.1] or Hirsch et al. [2012, Section 9.3].

Proposition 1 (Existence and uniqueness of gradient flows)**.** *Suppose the performative risk $PR(\cdot)$ is continuously differentiable, and its sublevel sets $\{x : PR(x) \leq c\}$ are compact for every $c \in \mathbb{R}$. Then for any initial condition $x_{PR}(0) = x_0$, there exists a unique solution to the differential equation in Equation* (4)*, defined for all $t \geq 0$.*

Next, we note that gradient flows have nice properties from the perspective of optimization. Namely: every isolated local minima is locally asymptotically stable, and we can provide sufficient conditions to characterize a subset of the region of convergence.

Proposition 2 (Convergence of gradient flows)**.** *Suppose the performative risk $PR(\cdot)$ is twice continuously differentiable, and x^* is an isolated local performative risk minimizer. Then x^* is a locally asymptotically stable equilibrium of Equation* (4)*. Furthermore, take any c such that $PR(x^*) \leq c$. Let $A \subseteq \{x : PR(x) \leq c\}$ denote the connected component of $\{x : PR(x) \leq c\}$ that contains x^*. If x^* is the only local performative minimizer in A, then all solutions with initial conditions in A converge to x^*.*

Proof. Since x^* is an isolated local minimizer and the performative risk is twice continuously differentiable, there exists a neighborhood $U \ni x^*$ such that $\nabla PR(\cdot)$ is non-zero for all $x \neq x^*$. By continuity, there exists some constant ϵ such that the connected component of $\{x : PR(x) \leq PR(x^*)+\epsilon\}$ containing x^* is contained in U. Since it is a sublevel set of $PR(\cdot)$ and $\mathcal{L}_{f_{PR}} PR(x) < 0$ on its boundary, it is positively invariant. Furthermore, since $\mathcal{L}_{f_{PR}}(x) < 0$ for all $x \neq x^*$ on this set, x^* is locally asymptotically stable by standard Lyapunov arguments (see, e.g. Khalil [2001, Section 4]). □

The sublevel sets of the performative risk are positively invariant with respect to the PRM gradient flow. Furthermore, because of the continuity of trajectories, each connected component will also be positively invariant. This, tandem with the fact that trajectories must either converge to a local minima or go off to infinity, also implies the previous proposition.

With minimal assumptions, isolated local performative risk minimizers are all locally attractive in the PRM gradient flow. In Section 5, we will view the PRM gradient flow as the nominal dynamics. From this perspective, we analyze the RRM flow as a perturbation from these nominal dynamics. To be able to do any perturbation-based analysis, we will need some stronger conditions on the convergence of the gradient flow associated with performative risk minimization. We note these assumptions here.

Assumption 1 (Sufficient curvature of the performative risk)**.** *Fix some isolated local performative risk minimizer x^*. We assume there exists positive constants c_1, c_2, c_3 and c_4 such that the following*

[3] As mentioned in the previous footnote, we augment x with p, the probability of seeing $\theta = 1$. Thus, although in both initialization, we see the classifier parameters are converging to the same point, with this augmentation, we can view these as two separate equilibria. However, for ease of presentation, we avoided cluttering notation with this augmentation.

holds in a neighborhood of x^*:

$$c_1|x - x^*|^2 \leq PR(x) - PR(x^*) \leq c_2|x - x^*|^2 \tag{7}$$

$$c_3|x - x^*| \leq |\nabla PR(x)| \leq c_4|x - x^*| \tag{8}$$

We will let r denote the radius of this neighborhood, so the above inequalities are valid on the set $\{x : |x - x^*| \leq r\}$.

Assumption 1 provides conditions on which $V(x) = PR(x) - PR(x^*)$ can be used as a Lyapunov function locally.

5 Analysis of repeated risk minimizing flow

In the previous section, we consider the PRM gradient flow and showed that the trajectories converge to local performative risk minimizers in very general settings. In this section, we will consider the RRM flow, defined by Equation (5). The RRM flow is not necessarily a gradient flow, and generally will not inherit the nice properties we saw in Section 4.

The following theorem provides conditions on the transient response and steady-state behavior of the RRM flow. Prior to T, the trajectories converge exponentially quickly. After T, we have an ultimate bound that holds.

Theorem 1 (Ultimate bounds for RRM flow)**.** *Fix any isolated performative risk minimizer x^* and suppose the conditions of Assumption 1 hold. Let $(c_i)_{i=1}^4$ denote the constants from Assumption 1 and $r > 0$ denote the radius where the inequalities are valid.*

Suppose that there exists positive constants $\epsilon < c_3^2/c_4$ and δ such that the following holds on $U = \{x : |x - x^*| \leq r\}$:

$$|\nabla_{x_2} R(x, x)| \leq \epsilon|x - x^*| + \delta \tag{9}$$

Additionally, suppose the initial condition satisfies:

$$|x_0 - x^*| \leq \sqrt{\frac{c_1}{c_2}} r$$

Take any $\theta \in (0, 1)$ such that:

$$\delta \leq \sqrt{\frac{c_2}{c_1}}(1 - \theta)r(c_3^2/c_4 - \epsilon)$$

Then, there exists a $T \geq 0$ such that:

- *For all $t \leq T$:*

$$|\varphi_{f_{RR}}(t; x_0) - x^*| \leq \sqrt{\frac{c_2}{c_1}} \exp(-t\theta(c_3^2 - c_4\epsilon)/2c_2)|x_0 - x^*|$$

- *For all $t \geq T$:*

$$|\varphi_{f_{RR}}(t; x_0) - x^*| \leq \sqrt{\frac{c_2}{c_1}} \frac{c_4\delta}{(1 - \theta)(c_3^2 - c_4\epsilon)}$$

Proof. Let $V(x) = R(x, x) - R(x^*, x^*)$. Note that $V(x) \geq 0$ on $U = \{x : |x - x^*| \leq r\}$ and $V(x) = 0$ if and only if $x = x^*$. Furthermore, note that $\frac{\partial V}{\partial x}(x) = [\nabla_{x_1} R(x, x) + \nabla_{x_2} R(x, x)]^\intercal$.

Consider the function $t \mapsto V(\varphi_{f_{RR}}(t; x_0))$ and its time derivative. Also, let $x_{RR}(t) = \varphi_{f_{RR}}(t; x_0)$. Taking the derivative along trajectories of the repeated risk minimization flow and using Equations (8) and (9):

$$\mathcal{L}_{f_{PR}+g} V = \frac{\partial V}{\partial x}(f_{PR}(x) + g) = -|\nabla_{x_1} R + \nabla_{x_2} R|^2 + \langle \nabla_{x_1} R + \nabla_{x_2} R, \nabla_{x_2} R \rangle \leq$$

$$-c_3^2|x_{RR} - x^*|^2 + c_4|x_{RR} - x^*||\nabla_{x_2} R| \leq -c_3^2|x_{RR} - x^*|^2 + c_4\epsilon|x_{RR} - x^*|^2 + c_4\delta|x_{RR} - x^*|$$

These inequalities are valid so long as $x_{RR}(t)$ stays within U, which we will ensure later in the proof. Note that ϵ is sufficiently small (by assumption) to ensure that $-c_3^2 + c_4\epsilon < 0$.

Let $\alpha := c_3^2 - c_4\epsilon > 0$. Take any $\theta \in (0, 1)$ and note that:

$$\mathcal{L}_{f_{PR}+g}V(x_{RR}) \leq -\theta\alpha|x_{RR} - x^*|^2 - (1-\theta)\alpha|x_{RR} - x^*|^2 + c_4\delta|x_{RR} - x^*|$$

Let $\mu(\theta) := c_4\delta/(1-\theta)\alpha$. If $|x_{RR} - x^*| \geq \mu(\theta)$, then:

$$\mathcal{L}_{f_{PR}+g}V(x_{RR}) \leq -\theta\alpha|x_{RR} - x^*|^2$$

Trajectories of Equation (5) has two stages: a transient due to its initial condition, and then an ultimate bound due to the perturbation. Let $T(\theta) = \inf\ \{t \geq 0 : |x_{RR}(t) - x^*| \leq \mu(\theta)\}$. Prior to $T(\theta)$, we have:

$$\frac{d}{dt}V(x_{RR}(t)) \leq -\theta\alpha|x_{RR}(t) - x^*|^2 \leq -\frac{\theta\alpha}{c_2}V(x_{RR}(t))$$

The latter follows from Equation (7). By the comparison principle (see, e.g. [Khalil, 2001, Lemma 3.4]), we have $V(x_{RR}(t)) \leq \exp(-t\theta\alpha/c_2)V(x_0)$. Again using Equation (7), this yields the following inequality, valid for all $t \leq T(\theta)$:

$$|x_{RR}(t) - x^*| \leq \sqrt{\frac{c_2}{c_1}}\exp(-t\theta\alpha/2c_2)|x_0 - x^*|$$

Note that this inequality also provides an upper bound on $T(\theta)$. Additionally, note that this implies the bound $|x_{RR}(t) - x^*| \leq r$, by our assumption on the initial condition. Prior to $T(\theta)$, our trajectory stays in U, where our inequalities are valid.

At time $T(\theta)$, we have $|x_{RR}(t) - x^*| \leq \mu(\theta)$. Note that this inequality implies $V(x_{RR}(t)) \leq c_2\mu^2(\theta)$. Since $\mathcal{L}_{f_{PR}+g}V < 0$ on the boundary of $\Omega(\theta) := \{x : V(x) \leq c_2\mu^2(\theta)\}$, we have that $\Omega(\theta)$ is a positively invariant set. So, for all $t \geq T(\theta)$, we have $x_{RR}(t) \in \Omega(\theta)$. Using Equation (7), we have the following for all $t \geq T(\theta)$:

$$|x_{RR}(t) - x^*| \leq \sqrt{\frac{c_2}{c_1}}\mu(\theta)$$

The condition on θ ensures that this quantity is bounded by r, and the trajectory stays in U for $t \geq T(\theta)$. This proves our desired result. □

Note that, in the special case where $\delta = 0$, we have that the RRM flow converges exponentially quickly to x^* locally. Similarly, in the special case where Assumption 1 holds everywhere (i.e. $r = \infty$), then there is only one minimizer x^*, and all initial conditions converge to a neighborhood of x^* exponentially fast.

Additionally, note that locally performatively stable points are equilibria of the RRM flow. This result provides constraints on where performatively stable points can be. Suppose again that Assumption 1 holds globally (i.e. $r = \infty$) and, consequently, there exists only one minimizer x^*. In this special case, Theorem 1 shows that all performatively stable points must be close to x^*, which is a continuous-time analog to Theorem 4.3 in Perdomo et al. [2020].

5.1 Performative alignment

From the previous analysis, we also identify conditions on the directions of the performative perturbations that are sufficient to show the convergence of Equation (5), the RRM flow, to performative risk minimizers.

Theorem 2 (Performative alignment). *Suppose x^* is a isolated local performative risk minimizer and the following holds for all x in a neighborhood of x^*:*

$$|\nabla_{x_2}R(x,x)|^2 \leq \langle -\nabla_{x_1}R(x,x), \nabla_{x_2}R(x,x)\rangle \tag{10}$$

Then x^* is a locally asymptotically stable equilibrium point of the RRM flow, given by Equation (5)*. Note that this does* ***not*** *require Assumption 1.*

Proof. Let $V(x) = PR(x) - PR(x^*)$. Since x^* is a locally asymptotically equilibria of the PRM flow, we have: $V(x^*) = 0$, $V(x) > 0$ for $x \neq 0$, and $\mathcal{L}_{f_{PR}}V(x) < 0$ for $x \neq 0$. The performative alignment condition ensures that $\mathcal{L}_{f_{PR}+g}V(x) < 0$ as well, and the desired result follows. □

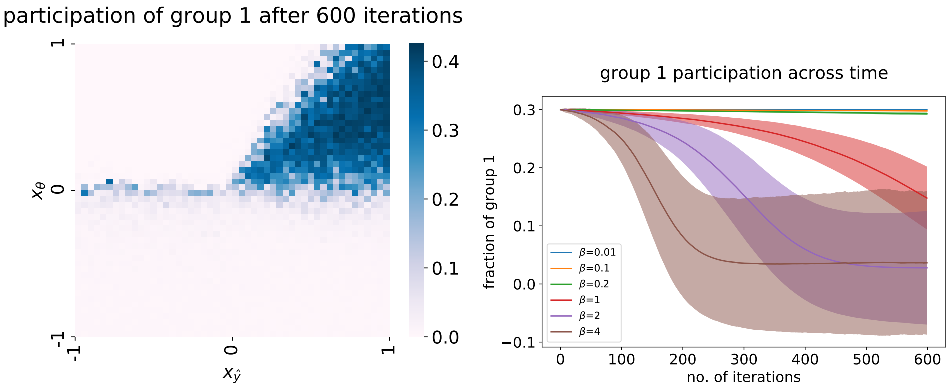

Figure 2: *Left:* The participation rate of group 1 after 600 iterations, based on the initialization of the weight vector $(x_\theta, x_{\hat{y}})$. All of these were initialized with $x_b = 0$. *Right:* The participation rate of group 1 across time for different scalings of the perturbation. Small perturbations (corresponding to smaller values of β) yield continued participation, whereas we fail to provide guarantees for larger perturbations (corresponding to larger values of β). This plot is shows the average and standard deviation across 100 trials.

We refer to Equation (10) as the **performative alignment** condition. This condition states that the performative perturbation never increases the performative risk, and the convergence of performative risk minimization is sufficient to guarantee convergence of repeated risk minimization. In other words, the perturbation is either sufficiently small or pointing in the correct direction to ensure that $PR(\cdot) - PR(x^*)$ can still act as a Lyapunov function.

Another perspective on performative alignment is to consider the performative risk as a bilinear form whose arguments are parameterized by x. In particular, consider the decoupled performative risk $R(\cdot, \cdot)$. Let $\ell_x := \ell(\cdot, x)$ and let μ_x denote the probability distribution associated with $\mathcal{D}(x)$. Then, we can write $R(x_1, x_2) = \langle \mu_{x_2}, \ell_{x_1} \rangle$. From this perspective, $R(\cdot, \cdot)$ is a bilinear form in ℓ_x and μ_x. As such, the performative alignment condition becomes a condition on the way in which ℓ and μ are *parameterized* by x.

6 Numerical results

In this section, we revisit the model introduced in Section 3.1. We again use the parameters: critical participation rate $p_c = 0.3$, saturation participation rate $\overline{p} = 0.5$, and observational noise $\sigma^2 = (0.8)^2$. The participation rate is initialized to $p = 0.3$. In lieu of explicitly calculating the expectation, we drew 100 data points at each iteration and used the empirical distribution to calculate update rules, as would typically be done in practice. All of the trials, including those in Section 3.1, were conducted with a constant step-size of $\alpha_k = 0.1$, and with the weight $\beta = 1$ for the update rule in Equation (6).

In Figure 2 *(Left)*, we visualize the participation rate of group 1 after 600 iterations, based on the initialization of the weight vector $(x_\theta, x_{\hat{y}})$. (In all cases, the bias term x_b was initialized to 0 and the initial participation was initialized to $p = 0.3$.) We can see the set of initial conditions in which group 1 continues to participate in the job application pool.

Next, we consider the effect of scaling the perturbation on the values of p across time. In this model, we can scale the perturbation by changing the weight parameter β in Equation (6). The bounds in Theorem 1 depend on the size of the perturbation; the region of attraction is larger for smaller perturbations. Figure 2 *(Right)* is conducted with the favored initialization $(x_\theta, x_{\hat{y}}, x_b) = (1, 1, 0)$, and with initial probability $p = 0.3$. The visualization shows the average value and standard deviation across 100 trials. We can see for small perturbations $\beta \in \{0.01, 0.1, 0.2\}$, the probability does not change significantly across time. For larger perturbations $\beta \in \{1, 2, 4\}$, the performative perturbation can be large enough to push us out of the region of attraction for the nominal dynamics.

7 Closing remarks

In this paper, we analyzed the problem of performative prediction in settings where multiple isolated equilibria may be of interest. We analyzed the gradient flow of performative risk minimization, and identified regions of attraction for various equilibria. We viewed repeated risk minimization flow as a perturbation of the PRM gradient flow. In particular, we used a Lyapunov function for the PRM gradient flow to analyze the trajectories of the RRM flow. We found conditions on which RRM flow will converge to the local PRM minimizers, and conditions on which they will converge to a neighborhood of PRM minimizers. Stochastic approximation results allowed us to state when repeated risk minimization will approximate the RRM flow studied.

These results provide a method to analyze the regions of attraction for various equilibria under repeated risk minimization. In real-world settings with decision-dependent distributions, we expect many situations where the initialization may have a significant outcome on the trajectories and final outcomes.

Acknowledgments and Disclosure of Funding

[TODO]

References

Emrah Akyol, Cedric Langbort, and Tamer Basar. Price of transparency in strategic machine learning. In *3rd Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2016.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. *arXiV*, 2020.

Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 547–555, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450308137. doi: 10.1145/2020408.2020495. URL https://doi.org/10.1145/2020408.2020495.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, EC '18, pages 55–70, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450358293. doi: 10.1145/3219166.3219193. URL https://doi.org/10.1145/3219166.3219193.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), 2018. doi: 10.1126/sciadv.aao5580. URL https://advances.sciencemag.org/content/4/1/eaao5580.

Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. *arXiV*, 2020.

Vikas Goel and Ignacio E. Grossmann. A stochastic programming approach to planning of offshore gas field developments under uncertainty in reserves. *Computers & Chemical Engineering*, 28(8): 1409–1429, 2004. ISSN 0098-1354. doi: https://doi.org/10.1016/j.compchemeng.2003.10.005. URL https://www.sciencedirect.com/science/article/pii/S0098135403002898.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16, pages 111–122, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340571. doi: 10.1145/2840728.2840730. URL https://doi.org/10.1145/2840728.2840730.

Lars Hellemo, Paul I. Barton, and Asgeir Tomasgard. Decision-dependent probabilities in stochastic programs with recourse. *Computational Management Science*, 15(3):369–395, 2018. doi: 10.1007/s10287-018-0330-0. URL https://doi.org/10.1007/s10287-018-0330-0.

Morris W. Hirsch, Stephen Smale, and Robert L. Devaney. *Differential Equations, Dynamical Systems, and an Introduction to Chaos*. Academic Press, 3rd edition, 2012.

Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 259–268, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287597. URL https://doi.org/10.1145/3287560.3287597.

T W Jonsbråten. Oil field optimization under price uncertainty. *Journal of the Operational Research Society*, 49(8):811–818, 1998. doi: 10.1057/palgrave.jors.2600562. URL https://doi.org/10.1057/palgrave.jors.2600562.

Tore W. Jonsbråten, Roger J-B Wets, and David L. Woodruff. A class of stochastic programs with decision dependent random elements. *Annals of Operations Research*, 82(0):83–106, 1998. doi: 10.1023/A:1018943626786. URL https://doi.org/10.1023/A:1018943626786.

Hassan K. Khalil. *Nonlinear Systems*. Pearson, 3rd edition, 2001.

Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4929–4939. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/33e75ff09dd601bbe69f351039152189-Paper.pdf.

John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6917–6926. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/miller20b.html.

John Miller, Juan C. Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. *arXiV*, 2021.

Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 230–239, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287576. URL https://doi.org/10.1145/3287560.3287576.

Srinivas Peeta, F. Sibel Salman, Dilek Gunnec, and Kannan Viswanath. Pre-disaster investment decisions for strengthening a highway network. *Computers & Operations Research*, 37(10):1708–1719, 2010. ISSN 0305-0548. doi: https://doi.org/10.1016/j.cor.2009.12.006. URL https://www.sciencedirect.com/science/article/pii/S0305054809003311.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7599–7609. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/perdomo20a.html.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]