Approximate Regions of Attraction in Learning with Decision-Dependent Distributions

Roy Dong
UIUC

Heling Zhang
UIUC

Lillian J. Ratliff
UW

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 situations where the initial conditions may have significant impact on the long-term behavior of the system. 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 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:

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

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:

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

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. Thus, when the decision-dependent distribution shift is not accounted for, the repeated gradient descent method yields the following update rule:

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

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 this paper, we shall analyze the steady-state behavior of the continuous-time flows corresponding to Equation (3):

$$\dot{x} = -\nabla_{x_1} R(x, x) \qquad (4)$$

The connections between the flow of Equation (4) and the repeated gradient descent method in Equation (3) can be drawn using results in stochastic approximation, i.e. the latter can be seen as a noisy forward Euler discretization of the former. For more details, we refer the reader to Borkar (2008).

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). 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 was first observed in Perdomo et al. (2020) under a different set of conditions. We note that Theorem 1 requires conditions on the curvature of the performative risk: the sublevel sets $\{x : PR(x) \leq c\}$ must grow in a precise fashion, such that upper and lower bounds on the performative risk $PR(x)$ imply upper and lower bounds on the norm of the argument x. Furthermore, the gradient of the performative risk must not vary too wildly around nearby points. This is formalized in Assumption 1. 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. This result does not require the strong curvature assumptions of Theorem 1.

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 provide motivating examples 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 Appendix B, and provide closing remarks in Section 6.

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

---

[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.

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.

Although our work still focuses on repeated risk minimization, it is worth noting that many other algorithms exist for learning with decision-dependent distributions. In Jagadeesan et al. (2022), the authors proposes the *performative confidence bounds algorithm* which uses tools from the bandit literature to explore the distribution map and find a near-optimal solution. In Izzo et al. (2021a), the authors proposes an algorithm called *performative gradient descent* (PerfGD), which guarantees to find the performatively optimal point when $\mathcal{D}(\cdot)$ satisfies certain parametric assumptions. Later in Izzo et al. (2021b), the author extends the results to settings where the distribution updates have internal states. In the same year, Li and Wai (2021) presents state-dependent stochastic approximation (SA) algorithm that works in similar settings. Of course, this paragraph is not a exhaustive treatment of various algorithms in similar settings.

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:

$$\begin{aligned}\dot{x}_{PR} &= -\nabla PR(x_{PR}) \\ &= -\nabla_{x_1} R(x_{PR}, x_{PR}) - \nabla_{x_2} R(x_{PR}, x_{PR}) \quad (5) \\ &=: f_{PR}(x_{PR})\end{aligned}$$

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 (5) 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 gradient descent (RGD) flow** as solutions to the differential equation:

$$\dot{x}_{RR} = -\nabla_{x_1} R(x_{RR}, x_{RR}) =: f_{RR}(x_{RR}) \quad (6)$$

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 RGD 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 RGD flow, which is the quantity of interest. In particular, we show ultimate bounds on the distance between the trajectories of RGD 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 $A \subset R^n$, 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. Additionally, we will use $\mathcal{W}_1(\cdot)$ to denote the **Wasserstein distance**, also known as the earth mover's distance.

3.1 EXAMPLES

Before we present our analysis of the PRM gradient flow and the RGD flow, we introduce some examples which motivate the study of performative risk in non-convex settings and multiple local equilibria.

3.1.1 Squared error loss and Bernoulli distributions

Consider the loss function $\ell(z, x) = \frac{1}{2}|z - x|^2$, where x is a scalar. Furthermore, suppose that the decision-dependent distribution $\mathcal{D}(x)$ is simply $Z = 1$ with probability $p(x)$ and $Z = 0$ with probability $1 - p(x)$, for some function $p(\cdot)$. In this case, the decoupled performative risk is given by:

$$R(x_1, x_2) = p(x_2) \left[\frac{1}{2}|1 - x_1|^2 \right] + (1 - p(x_2)) \left[\frac{1}{2}|x_1|^2 \right]$$
$$= \frac{1}{2}[x_1^2 + p(x_2)(1 - 2x_1)]$$
(7)

We will analyze this model in two ways. First, we will consider general $p(\cdot)$, and, fixing the loss function $\ell(\cdot)$, identify a class of decision-dependent distribution shifts $p(\cdot)$ which can still ensure convergence to performative risk minimizers, using Theorem 2. Second, we will consider a concrete example for $p(\cdot)$, and demonstrate how to apply Theorem 1 to understand the regions of convergence.

As our concrete example of $p(\cdot)$, consider the following function as a candidate for $p(\cdot)$:

$$\varphi(x) = \begin{cases} \exp\left(1 + \frac{-1}{1-(x-1)^2}\right) & \text{if } x \in (0, 1) \\ 0 & \text{if } x \leq 0 \\ 1 & \text{if } x \geq 1 \end{cases} \quad (8)$$

This function is chosen because $\varphi(x) = 1$ for $x \geq 1$, $\varphi(x) = 0$ for $x \leq 0$, and it is continuously differentiable. The derivative is:

$$\varphi'(x) = \begin{cases} \varphi(x) \frac{2(1-x)}{(1-(x-1)^2)^2} & \text{if } x \in (0, 1) \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

$\varphi(\cdot)$ and its derivative is visualized in Figure 1(a). Since $p(0) = 0$ and $p(1) = 1$ for this choice of $p(\cdot)$, Equation (7) directly implies that there are two performative risk minimizers: $x = 0$ and $x = 1$. Similarly, we can see that these points are performatively stable as well. The corresponding performative risk and gradients are visualized in Figure 1(b)–(c).

3.1.2 Classification of adversarial agents

As we've mentioned, when data-driven algorithms are deployed in real-world settings, it often caused a drift in the distribution of the data. One source of such drift is the behavior of adversarial agents. Here, we consider a simple classification problem of potentially adversarial agents.

Suppose that each agent are defined by a feature $z \in R^d$ and a binary label $y \in \{-1, 1\}$, and suppose that they are drawn i.i.d from a distribution $\mathcal{D}_{Z,Y}$. The task of the classifier is to correctly predict their labels based on their features. For simplicity, we further assume that we are using a linear classifier defined by

$$\hat{y}_x = sign(\langle x, z \rangle),$$

where $x \in R^d$ is the parameter of the classifier. We also assume that the learner is using a logistic loss function:

$$\ell(x, y, z) = \log(1 + \exp(-y\langle x, z \rangle)).$$

So far, we've essentially described a ordinary binary classification problem with linear classifiers. We further assume that once the classifier is deployed, the agents will potentially alter their features to induce false predictions. Formally, we assume that if an agent is adversarial, it will produce a fake feature $\hat{z}$ such that

$$\hat{z}(x, y, z) \in \operatorname*{argmin}_{z'} \left\{ -y\langle x, z' \rangle + \frac{k}{2} \|z' - z\|_2^2 \right\},$$

and that agents are adversarial with probability

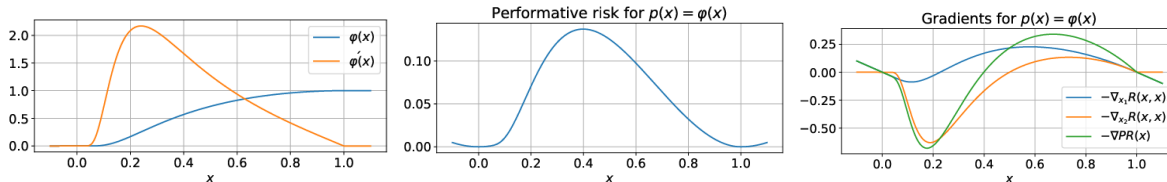$$p_{adv}(x, y, z) = e^{-\lambda_1 \|\hat{z}(x,y,z) - z\|_2^2}.$$

Figure 1: As an illustrative example, we consider a setting where with the squared error is used as the loss function, and the decision-dependent distribution shift modifies the parameters of a Bernoulli distribution, as discussed in Section 3.1.1. (a) A visualization of an example decision-dependent distribution shift $\varphi(x)$, as defined in Equation (8), and its derivative, as derived in Equation (9). (b) The performative risk with $\Pr(Z = 1) = p(x) = \varphi(x)$. (c) The corresponding gradients for $p(x) = \varphi(x)$.

Essentially, we are implicitly describing a 'cost' of adversarial behavior, which that is proportional to the squared l_2 distance between true and fake features. If the cost is too high, an agent will be more likely to abandon adversarial behavior, hence a lower p_{adv}.

This adversarial behavior can be thought of a distribution shift caused by deploying our classifier, and the performative risk is then given by

$$PR(x) = E_{(\hat{Z},Y)\sim\mathcal{D}(x)}\left[\ell(x, Y, \hat{Z})\right]$$
$$= E_{(Z,Y)\sim\mathcal{D}_{Z,Y}}\left[(1 - p_{adv}(x, Y, Z))\ell(x, Y, Z)\right.$$
$$\left. + p_{adv}(x, Y, Z)\ell(x, Y, \hat{Z}(x, Y, Z)).$$

Our goal here is to find a classifier that minimizes $PR(x)$ without knowing the behavior pattern of the agents. That is, we are completely unaware of the dependencies of the performative risk on x_2.

To illustrate, consider the simplest case where $d = 1$ (i.e. $z, x \in R$). Specifically, let $\mathcal{D}_{Z,Y} = \mathcal{N}(y, 1)$. The corresponding performative risk and gradients are illustrated in Figure 2(d) and (e).

In both these examples, there are multiple performative risk minimizers and performatively stable points. Performative risk minimization and repeated gradient descent can converge to different steady-state results, and it is of interest which initializations will converge to which equilibria under both dynamics. In the sequel, we shall demonstrate how different functions $p(\cdot)$ can lead to different steady-state outcomes, as well as how our theoretical results can provide conditions on $p(\cdot)$ such that we achieve convergence to performative risk minimizers, even when performing repeated approximate risk minimization.

4 ANALYSIS OF PERFORMATIVE RISK MINIMIZING GRADIENT FLOW

In this section, we consider PRM gradient flow, defined by Equation (5). 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 (5) 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 $\{x : PR(x) \le c\}$ are compact for every $c \in R$. Then for any initial condition $x_{PR}(0) = x_0$, there exists a unique solution to the differential equation in Equation (5), defined for all $t \ge 0$.*

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 (5). Furthermore, take any c such that $PR(x^*) \le c$. Let $A \subseteq \{x : PR(x) \le c\}$ denote the connected component of $\{x : PR(x) \le c\}$ 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 $x \ne x^*$. By continuity, there exists some constant ϵ such that the connected component of $\{x : PR(x) \le PR(x^*) + \epsilon\}$ 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 $x \ne x^*$ on this set, x^* is locally asymptotically stable by standard Lyapunov argu-
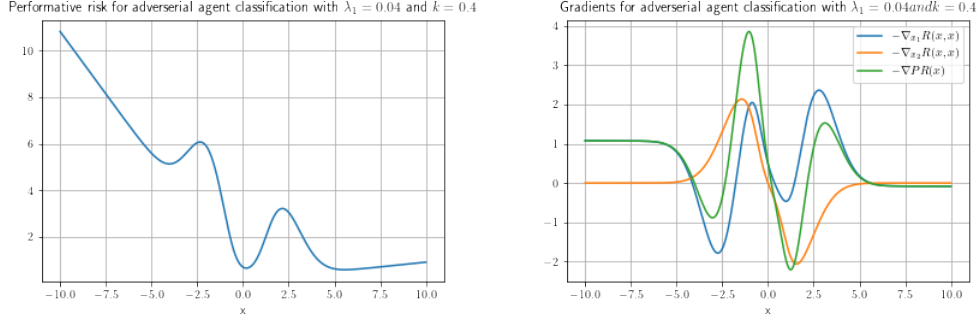
Figure 2: This is a 1-dimensional illustration of the setting described in Section 3.1.2. When $(z, y) \sim \mathcal{N}(y, 1)$, $\lambda_1 = 0.04$ and $k = 0.4$, the corresponding performative risk is visualized in (d), and the corresponding gradients are visualized in (e).

ments (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, in 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 RGD 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 PR). *Fix some isolated local performative risk minimizer x^*. We assume there exists positive constants c_1, c_2, c_3, c_4 and δ such that the following holds in a neighborhood of x^*:*

$$c_1|x - x^*|^2 \le PR(x) - PR(x^*) \le c_2|x - x^*|^2 \quad (10)$$

$$c_3|x - x^*| - \delta \le |\nabla PR(x)| \le c_4|x - x^*| + \delta \quad (11)$$

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

Assumption 1 provides conditions on which $V(x) = PR(x) - PR(x^*)$ can be used as a Lyapunov function locally. Next, we provide conditions directly on the loss $\ell(\cdot)$ and the decision-dependent distribution shift $\mathcal{D}(\cdot)$ which can ensure that Assumption 1 holds, or approximately holds. First, we provide sufficient conditions for the bounds in Equation (10).

Proposition 3 (Performative risk bounds). *Let x^* be a performative risk minimizer and fix any x. If:*

1. $\ell(\cdot, x)$ *is L_1 Lipschitz continuous*

2. $\mathcal{W}_1(\mathcal{D}(x), \mathcal{D}(x^*)) \le L_2|x - x^*|^2$

3. $\ell(z, \cdot)$ *is m-strongly convex and L_3-smooth for every z*

Then: $(m/2 - L_1L_2)|x - x^*|^2 \le PR(x) - PR(x^*) \le (L_1L_2 + L_3/2)|x - x^*|^2$.

Proof. First, we can break up the performative risk into two parts: $PR(x) - PR(x^*) = R(x, x) - R(x^*, x^*) = [R(x, x) - R(x, x^*)] + [R(x, x^*) - R(x^*, x^*)]$. Note that $R(x, x) - R(x, x^*) = E_{Z \sim \mathcal{D}(x)}[\ell(Z, x)] - E_{Z \sim \mathcal{D}(x^*)}[\ell(Z, x)]$. Conditions (1) and (2), along with Kantorovich-Rubenstein duality (Villani, 2003), implies this quantity is bounded in absolute value: $|R(x, x) - R(x, x^*)| \le L_1L_2|x - x^*|^2$. On the other hand, $R(x, x^*) - R(x^*, x^*) = E_{Z \sim \mathcal{D}(x^*)}[\ell(Z, x) - \ell(Z, x^*)]$. By convexity and L_3-smoothness, $\ell(z, x) - \ell(z, x^*) \le \langle \nabla_x\ell(z, x^*), x - x^* \rangle + \frac{L_3}{2}|x - x^*|^2$ for any z; taking the expectation and noting that $\nabla PR(x^*) = 0$, we have $R(x, x^*) - R(x^*, x^*) \le \frac{L_3}{2}|x - x^*|^2$. In the other direction, using strong convexity and similar arguments, we get: $R(x, x^*) - R(x^*, x^*) \ge \frac{m}{2}|x - x^*|^2$. Combining these results yields the desired results. □

Note that Condition (2) in Proposition 3 is a variation on the typical ϵ-sensitivity definition. Recall that ϵ-sensitivity states that for any x and y, $\mathcal{W}_1(\mathcal{D}(x), \mathcal{D}(y)) \le \epsilon|x - y|$ (Perdomo et al., 2020). In contrast, Condition (2) only requires this condition to hold around the point x^*, but requires a stricter bound for x close to x^*. This bound is also more lax than ϵ-sensitivity farther away from x^*.

Next, we provide sufficient conditions for a bound on the absolute value of the gradient of the performative risk. This does not exactly recover Assumption 1, as there are additive constants on the upper and lower bounds which do not scale with $|x - x^*|$. However, these additive constants will

be small for decision-dependent distribution shifts $\mathcal{D}(\cdot)$ with sufficiently small sensitivity parameter ϵ.

Proposition 4 (Gradient bounds of the performative risk). *Let x^* be a performative risk minimizer and fix any x. If:*

1. *$\ell(\cdot, x)$ and $\ell(\cdot, x^*)$ are both L_1 Lipschitz continuous*

2. *$\ell(z, \cdot)$ is m-strongly convex and L_3-smooth for every z*

3. *$\mathcal{D}(\cdot)$ is ϵ-sensitive, i.e. $\mathcal{W}_1(\mathcal{D}(x), \mathcal{D}(y)) \leq \epsilon|x - y|$*

4. *$\nabla_x \ell(\cdot, x)$ is L_4 Lipschitz continuous*

Then: $(m - \epsilon L_4)|x - x^*| - 2\epsilon L_1 \leq |\nabla PR(x)| \leq (L_3 + \epsilon L_4)|x - x^*| + 2\epsilon L_1$.

Proof. Similar to the previous proposition, we break apart this gradient. Note that $\nabla PR(x^*) = 0$, so: $|\nabla PR(x)| = |\nabla PR(x) - \nabla PR(x^*)| = |\nabla_{x_1} R(x, x) - \nabla_{x_1} R(x^*, x^*) + \nabla_{x_2} R(x, x) - \nabla_{x_2} R(x^*, x^*)|$. For the ∇_{x_1} terms, we have: $m|x - x^*| \leq |\nabla_{x_1} R(x, x^*) - \nabla_{x_1} R(x^*, x^*)| \leq L_3|x - x^*|$ by standard convexity arguments, and $|\nabla_{x_1} R(x, x) - \nabla_{x_1} R(x, x^*)| \leq \epsilon L_4|x - x^*|$ by the same Kantorovich-Rubenstein duality argument as the previous proposition. For the ∇_{x_2} terms, note that the mapping $x_2 \mapsto R(x, x_2)$ is ϵL_1 Lipschitz continuous. Thus, $|\nabla_{x_2} R(x, x)| \leq \epsilon L_1$ and similarly $|\nabla_{x_2} R(x^*, x^*)|$. Combining these inequalities yields the desired result. $\square$

Depending on the situation, we may be able to directly verify Assumption 1, although, for more complex settings, this is likely to be very difficult. Propositions 3 and 4 provide a set of sufficient conditions for this assumption to approximately hold, but checking the conditions on the decision-dependent distribution shift $\mathcal{D}(\cdot)$ may be difficult in practice as well. This is one limitation of this current work, and we believe it is an interesting future research direction to identify conditions which are easy to verify, even in settings with limited information about the distribution shift itself.

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 RGD flow, defined by Equation (6). The RGD 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 RGD flow. Prior to T, the trajectories converge exponentially quickly. After T, we have an ultimate bound that holds.

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

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

$$|\nabla_{x_2} R(x, x)| \leq \epsilon|x - x^*| + \delta \qquad (12)$$

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:

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

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 $t \geq T$:*

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

Proof. See Appendix A. $\square$

Note that, in the special case where $\delta = \lambda = 0$, we have that the RGD 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. In addition, if every condition in Assumption 1 holds with equality, the bounds in Theorem 1 also hold with equality. Please see Appendix A for an example where this occurs.

Additionally, note that locally performatively stable points are equilibria of the RGD flow. This result provides constraints on where performatively stable points can be located. Consider the special case where 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^*. The phenomena that, under certain conditions, performatively stable points are near performative risk minimizers, was first noted in Perdomo et al. (2020). Our results here

provide another set of conditions under which the same result holds.

Furthermore, in the presence of Proposition 3 and 4, Theorem 1 can be restated in terms of $\ell(\cdot)$ and $\mathcal{D}(\cdot)$ as follows:

Corollary 1. *Let x^* be a performative risk minimizer and fix any x. If:*

1. *$\ell(\cdot, x)$ and $\ell(\cdot, x^*)$ are both L_1 Lipschitz continuous*

2. *$\ell(z, \cdot)$ is m-strongly convex and L_3-smooth for every z*

3. *$\mathcal{D}(\cdot)$ is ϵ-sensitive, i.e. $\mathcal{W}_1(\mathcal{D}(x), \mathcal{D}(y)) \leq \epsilon |x - y|$*

4. *$\mathcal{W}_1(\mathcal{D}(x), \mathcal{D}(x^*)) \leq L_2 |x - x^*|^2$*

5. *$\nabla_x \ell(\cdot, x)$ is L_4 Lipschitz continuous*

Suppose the initial condition satisfies:

$$|x_0 - x^*| \leq \sqrt{\frac{m - 2L_1 L_2}{L_3 + 2L_1 L_2}} r$$

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

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

$$|\varphi_{f_{RR}}(t; x_0) - x^*| \leq$$

$$\sqrt{\frac{L_3 + 2L_1 L_2}{m - 2L_1 L_2}} \exp\left(\frac{-t\theta(m - \epsilon L_4)^2}{L_3 + 2L_1 L_2}\right) |x_0 - x^*|$$

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

$$|\varphi_{f_{RR}}(t; x_0) - x^*| \leq \sqrt{\frac{L_3 + 2L_1 L_2}{m - 2L_1 L_2}}$$

$$\cdot \max\left\{\frac{4\epsilon L_1(L_3 + 2m - \epsilon L_4)}{(m - \epsilon L_4)^2}, \frac{2\epsilon L_1}{m - \epsilon L_4}\right\}.$$

Proof. This follows immediately by combining Theorem 1 with Propositions 3 and 4. $\square$

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 (6), the RGD 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^*:*

$$|\nabla_{x_2} R(x, x)|^2 \leq \langle -\nabla_{x_1} R(x, x), \nabla_{x_2} R(x, x) \rangle \quad (13)$$

*Then x^* is a locally asymptotically stable equilibrium point of the RGD flow, given by Equation (6). 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. $\square$

We refer to Equation (13) 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 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.

In Appendix B.2, we apply Theorem 2 to the example outlined in Section 3.1.1. It provides insight into one of the ways to use Theorem 2: when we fix a loss $\ell(\cdot)$, we can view the performative alignment condition as specifying a class of decision-dependent distribution shifts which do not hamper the convergence of RGD to performative risk minimizers.

6 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 gradient descent 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 RGD flow. We found conditions on which RGD flow will converge to the local PRM minimizers, and conditions on which they will converge to a neighborhood of PRM minimizers.

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.

References

Akyol, E., Langbort, C., and Basar, T. (2016). Price of transparency in strategic machine learning. In *3rd Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Borkar, V. S. (2008). *Stochastic Approximation: A Dynamical Systems Viewpoint*. Hindustan Book Agency.

Brown, G., Hod, S., and Kalemaj, I. (2020). Performative prediction in a stateful world. *arXiV*.

Brückner, M. and Scheffer, T. (2011). 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. Association for Computing Machinery.

Dong, J., Roth, A., Schutzman, Z., Waggoner, B., and Wu, Z. S. (2018). 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. Association for Computing Machinery.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1).

Drusvyatskiy, D. and Xiao, L. (2020). Stochastic optimization with decision-dependent distributions. *arXiV*.

Goel, V. and Grossmann, I. E. (2004). A stochastic programming approach to planning of offshore gas field developments under uncertainty in reserves. *Computers & Chemical Engineering*, 28(8):1409–1429.

Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. (2016). Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16, pages 111–122, New York, NY, USA. Association for Computing Machinery.

Hellemo, L., Barton, P. I., and Tomasgard, A. (2018). Decision-dependent probabilities in stochastic programs with recourse. *Computational Management Science*, 15(3):369–395.

Hirsch, M. W., Smale, S., and Devaney, R. L. (2012). *Differential Equations, Dynamical Systems, and an Introduction to Chaos*. Academic Press, 3rd edition.

Hu, L., Immorlica, N., and Vaughan, J. W. (2019). 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. Association for Computing Machinery.

Izzo, Z., Ying, L., and Zou, J. Y. (2021a). How to learn when data reacts to your model: Performative gradient descent. In *International Conference on Machine Learning*.

Izzo, Z., Zou, J. Y., and Ying, L. (2021b). How to learn when data gradually reacts to your model. *ArXiv*, abs/2112.07042.

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

Jonsbråten, T. W. (1998). Oil field optimization under price uncertainty. *Journal of the Operational Research Society*, 49(8):811–818.

Jonsbråten, T. W., Wets, R. J.-B., and Woodruff, D. L. (1998). A class of stochastic programs with decision dependent random elements. *Annals of Operations Research*, 82(0):83–106.

Khalil, H. K. (2001). *Nonlinear Systems*. Pearson, 3rd edition.

Li, Q. and Wai, H.-T. (2021). State dependent performative prediction with stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*.

Mendler-Dünner, C., Perdomo, J., Zrnic, T., and Hardt, M. (2020). Stochastic optimization for performative prediction. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4929–4939. Curran Associates, Inc.

Miller, J., Milli, S., and Hardt, M. (2020). Strategic classification is causal modeling in disguise. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6917–6926. PMLR.

Miller, J., Perdomo, J. C., and Zrnic, T. (2021). Outside the echo chamber: Optimizing the performative risk. *arXiV*.

Milli, S., Miller, J., Dragan, A. D., and Hardt, M. (2019). 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. Association for Computing Machinery.

Peeta, S., Sibel Salman, F., Gunnec, D., and Viswanath, K. (2010). Pre-disaster investment decisions for strengthening a highway network. *Computers & Operations Research*, 37(10):1708–1719.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. (2020). Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7599–7609. PMLR.

Villani, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society.

A PROOF OF THEOREM 1

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 $\frac{\partial V}{\partial x}(x) = [\nabla_{x_1} R(x, x) + \nabla_{x_2} R(x, x)]$.

Consider the function $t \mapsto V(\varphi_{f_{RR}}(t; x_0))$ and its time derivative. Also, let $x_{RR}(t) = \varphi_{f_{RR}}(t; x_0)$. When $|x_{RR} - x^*| \geq \delta/c_3$, taking the derivative along trajectories of the repeated risk minimization flow and using Equations (11) and (12):

$$
\begin{aligned}
\mathcal{L}_{f_{PR}+g} V = \frac{\partial V}{\partial x}(f_{PR}(x) + g) &= -|\nabla_{x_1} R + \nabla_{x_2} R|^2 + \langle \nabla_{x_1} R + \nabla_{x_2} R, \nabla_{x_2} R \rangle \\
&\leq -(c_3|x_{RR} - x^*| - \delta)^2 + (c_4|x_{RR} - x^*| + \delta)|\nabla_{x_2} R| \\
&\leq -(c_3|x_{RR} - x^*| - \delta)^2 + (c_4|x_{RR} - x^*| + \delta)(\epsilon|x_{RR} - x^*| + \delta) \\
&= -(c_4\epsilon - c_3^2)|x_{RR} - x^*|^2 + (c_4\delta + 2c_3\delta + \delta\epsilon)|x_{RR} - x^*|
\end{aligned}
$$

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:

$$
\mathcal{L}_{f_{PR}+g} V(x_{RR}) \leq -\theta\alpha|x_{RR} - x^*|^2 - (1 - \theta)\alpha|x_{RR} - x^*|^2 + (c_4\delta + 2c_3\delta + \delta\epsilon)|x_{RR} - x^*|
$$

Now, let

$$
\mu(\theta) := \max\left\{ \frac{(c_4\delta + 2c_3\delta + \delta\epsilon)}{\alpha(1 - \theta)}, \frac{\delta}{c_3} \right\}.
$$

If $|x_{RR} - x^*| \geq \mu(\theta)$, then:

$$
-\theta\alpha|x_{RR} - x^*|^2 - (1 - \theta)\alpha|x_{RR} - x^*|^2 + (c_4\delta + 2c_3\delta + \delta\epsilon)|x_{RR} - x^*| \leq 0,
$$

and thus

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

Trajectories of Equation (6) 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 (10). 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 (10), 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 (10), 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.

Additionally, we can show that this bound is tight by considering the following example. Suppose $\mathcal{D}(x_2)$ is the point mass distribution (i.e. $p(z) = \delta(z - x_2)$) and $l(z, x_1) = 1/2|z|^2 + 1/2|x_1|^2$. Then the performative risk is given by $R(x_1, x_2) = 1/2|x_1|^2 + 1/2|x_2|^2$. It follows that x^*0 is the performative risk minimizer. Following the arguments in Appendix A, one would find that the dynamics of $V(x) = R(x, x) - R(x^*, x^*)$ follows $\frac{d}{dt} V(x(t)) = -2|x(t) - x^*|^2 = -2V(x(t))$, which yields $|x(t) - x^*| = \exp(-2t)|x_0 - x^*|$.
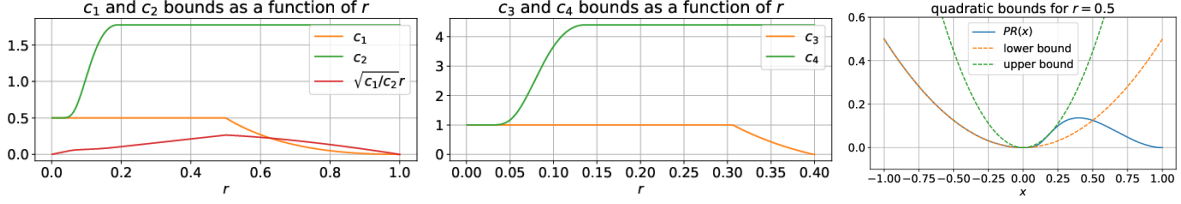
Figure 3: We verify that the performative risk bounds in Assumption 1 are satisfied in the example discussed in Section 3.1. (a) As a function of r (the radius of the domain where the inequalities hold), we show the tightest constants c_1 and c_2 for the bound. We also plot $\sqrt{c_1/c_2}r$, which is the radius of a neighborhood of $x = 0$ to which Theorem 1 can be applied. (b) As a function of r, we show the tightest constants for c_3 and c_4. (c) Choosing the c_1 and c_2 constants for $r = 0.5$, we visualize how the quadratic bounds hold for the performative risk locally.

B NUMERICAL EXAMPLES

In this section, we revisit the models introduced in Section 3.1. We demonstrate how the results of Sections 4 and 5 can be applied. First, we show that the example satisfies Assumption 1 and we calculate its corresponding constants. Second, we apply Theorem 1 and show the theoretical convergence rates match simulated trajectories. Finally, we also apply Theorem 2 to the example from Section 3.1.1 and characterize the class of distribution shifts satisfy the performative alignment condition.

B.1 Checking the curvature of the performative risk and region of convergence

Recall the example from Section 3.1.1, where x was a scalar, the loss function was the squared error, and the decision-dependent distribution was a Bernoulli random variable whose distribution was determined by $p(\cdot)$. In this section, we consider the specific decision-dependent distribution shift $p = \varphi$, which is defined in Equation (8).

When we consider this example, we can see that the bounds on Assumption 1 cannot hold globally, which matches our previous observation that there are multiple isolated performative risk minimizers. However, these bounds may hold locally: we can view the constants $(c_i)_{i=1}^4$ from Assumption 1 as a function of the size of the domain r.

For concreteness, let us focus on the equilibrium point $x = 0$. Recall that Assumption 1 must hold locally, on the domain $\{x : |x - x^*| \leq r\}$. As we increase r, the constants will worsen; we visualize this in Figure 3(a)–(b). Note that these bounds only have to hold locally around the equilibria, as visualized in Figure 3(c). Furthermore, the gradient bounds in Assumption 1 cannot hold beyond $r > 0.40$, since $\nabla PR(x) = 0$ at that point.

Recall that the convergence results of Theorem 1 can only apply to all initial conditions satisfying $|x_0 - x^*| < \sqrt{c_1/c_2}r$; we visualize this as well in Figure 3(a). On the set $(0, 0.40]$, we can see the quantity $\sqrt{c_1/c_2}r$ is the largest at $r = 0.4$, with constants $c_1 = 0.50$ and $c_2 = 1.78$. Thus, around the equilibrium $x = 0$, the theorem can be applied to all points in the set $\{x : |x| \leq 0.21\}$, with $\delta = 0$. Thus, our theorem shows that all points in this neighborhood of $x = 0$ will converge. This under-approximates the true region of attraction, which we numerically saw to be $\{x : x < 0.23\}$.

B.2 Performative alignment with squared error and Bernoulli distributions

We again consider the example from Section 3.1.1. However, in this section, we consider a general decision-dependent distribution shift $p(\cdot)$. We suppose that $p(0) = 0$ and $p(1) = 1$, so we have two performative risk minimizers as in our previous example. We have $\nabla_{x_1} R(x, x) = x - p(x)$ and $\nabla_{x_2} R(x, x) = (1/2 - x)p'(x)$. The performative alignment condition becomes:

$$|1/2 - x|^2 |p'(x)|^2 \leq (p(x) - x)(1/2 - x)p'(x) \tag{14}$$

Theorem 2 states that if this condition holds for all $x \in (0, c)$, then any initial conditions $x_0 \in (0, c)$ will converge to $x = 0$. Similarly, if this condition holds for all $x \in (c, 1)$, then all initial conditions in $(c, 1)$ will converge to $x = 1$. Theorem 2 also implies that this condition cannot be satisfied for all $x \in (0, 1)$, as then these initial conditions would converge to *both* $x = 0$ and $x = 1$.

If we suppose that $p(\cdot)$ is monotonic on $(0, 1)$, i.e. $p'(x) \geq 0$, we can also interpret the performative alignment condition as follows. For $x \in (1/2, 1)$, the performative alignment condition becomes $p(x) - x \geq (1/2 - x)p'(x)$. In this regime,

$(1/2 - x)p'(x) \leq 0$. In this setting, if $p(x) - x$ is too negative, the RGD flow will push x away from the nearby minimizer $x = 1$. Similarly, for $x \in (0, 1/2)$, the condition becomes $p(x) - x \leq (1/2 - x)p'(x)$. In this regime, $(1/2 - x)p'(x) \geq 0$, and the condition states that $p(x) - x$ cannot be too large, or the RGD flow will push x away from the minimizer $x = 0$.

In this section, we used Theorem 2 to identify conditions on the decision-dependent distribution shift $p(\cdot)$ which ensure that the performative risk does not increase even when the dynamics follow repeated gradient descent. For this example, the condition is that p satisfies Equation (14) for all $x \in (0, c)$. More generally, the performative alignment condition allow us to specify a class of distribution shifts which behave well with respect to performative risk minimization.