Optimizing the Performative Risk under Weak Convexity Assumptions

author names withheld

Under Review for OPT 2022

Abstract

In performative prediction, a predictive model impacts the distribution that generates future data, a phenomenon that is being ignored in classical supervised learning. In this closed-loop setting, the natural measure of performance, named performative risk (PR), captures the expected loss incurred by a predictive model *after* deployment. The core difficulty of using the performative risk as an optimization objective is that the data distribution itself depends on the model parameters. This dependence is governed by the environment and not under the control of the learner. As a consequence, even the choice of a convex loss function can result in a highly non-convex PR minimization problem. Prior work has identified a pair of general conditions on the loss and the mapping from model parameters to distributions that implies convexity of the performative risk. In this paper, we relax these assumptions and focus on obtaining weaker notions of convexity, without sacrificing the amenability of the PR minimization problem for iterative optimization methods.

1. Introduction

Predictions in the real world are often performative. This means predictions because they are made public, or because they inform down-stream actions, influence the distribution of the features or the target variable they aim to predict. Thus, shifting the data distribution the model has been trained on. Prominent examples include stock price prediction, credit scoring, and predictive policing.

As a result the performance of a model should not be measured with respect to a fixed underlying distribution, as in classical supervised learning, but instead on the very distribution the model induces. This has been conceptualized by [29] through the framework of *performative prediction*.

This dynamic setting gives rise to interesting and new optimization challenges. The objective function in performative prediction is the risk incurred by the learner on the distribution induced after deploying the model:

$$\mathrm{PR}(\theta) := \mathbb{E}_{z\sim\mathcal{D}(\theta)}\ell(z;\theta), \theta \in \Theta \tag{1}$$

where Θ is a closed and convex set in $\mathbb{R}^d$. PR is referred to as the performative risk and measures the loss of a predictor θ on the distribution $\mathcal{D}(\theta)$ resulting from its deployment. We focus on finding minima of the performative risk (1), known as performatively optimal points. The dependence of the objective on the model parameters through the loss as well as the data distribution makes this optimization problem particularly interesting to study.

Optimizing the performative risk directly is challenging because the structure of the objective $\mathrm{PR}(\theta) = \mathbb{E}_{z\sim\mathcal{D}(\theta)}\ell(z;\theta)$ is not under the control of the learner and can be highly non-convex. Thus,

© .

for tackling performative risk minimization by building on classical tools from the optimization literature we need to better understand the structure of the objective function, and how properties of the distribution map and the loss interact.

As an important first step, Miller et al. [25] identified a natural set of properties of the loss function and model-induced distribution shift under which the performative risk is convex. In particular, they showed that linear dependence of the distribution map on the model parameters θ, together with strong convexity of the loss function ℓ implies global convexity of PR. In this work, we are interested in studying problems beyond this specific setting and we intend to address the following question:

How and under what conditions can we optimize the performative risk?

In particular, we study how to optimize the performative risk through first-order information and analyze the suboptimality gap of performative risks. Noticing the fact that convexity can be relaxed to weaker conditions while same convergence guarantees are preserved by gradient-based optimizer, in Section 3, we show weaker conditions on DPR still imply desirable local properties of PR. The properties enjoy generality since they do not rely on global structural restrictions. In Section 4, we discuss the information about θ_{PO} we have when starting from a stable point. The underlying idea is to make use of previous retraining methods [23, 29] converging to θ_{PS} while the ultimate goal is still reaching θ_{PO}.

2. Background

Performativity is a phenomenon that is widely recognized in finance [6], economics [5, 7] and other social sciences [31]. Following the seminal work by [29] performativity has increasingly gained attention by the machine learning community. Performative distribution shifts constitute an instance of a more general distribution shift problem, where the shift is endogenous to the problem and entirely caused by the deployed predictive model. In other words, the deployment of a predictive model represents an intervention to the population [24]. The resulting causal feedback effects invalidate the standard assumption of supervised learning where there is a static and model-independent data distribution describing the population, surfacing as distribution shift.

2.1. Performative prediction framework

The framework of performative prediction proposed by [29] conceptualizes the role of performative feedback in supervised learning. They introduce the conceptual device of a distribution map $\mathcal{D}(\theta)$, mapping model parameters to a distribution over instances. Given the non-static nature of the resulting risk minimization problem, the quality of a predictive model parameterized by θ is measured by its risk on the distribution that it induces, denoted by

$$\mathrm{PR}(\theta) := \mathbb{E}_{z\sim\mathcal{D}(\theta)}\ell(z;\theta),$$

where PR stands for performative risk. The main challenge of performative risk minimization from an optimization perspective is that the dependence of the performative risk on the model parameters is two-fold: it appears both in the distribution $\mathcal{D}$ and the loss ℓ. To decouple these two dependencies, let us define the *decoupled performative risk*

$$\mathrm{DPR}(\theta_1,\theta_2) = \mathbb{E}_{z\sim\mathcal{D}(\theta_1)}\ell(z;\theta_2)$$

to denote the loss of θ_2 measured over the distribution $\mathcal{D}(\theta_1)$. According to this definition we have $\mathrm{PR}(\theta) = \mathrm{DPR}(\theta, \theta)$.

2.2. Main concepts

We first give an overview of two existing concepts characterizing whether θ is a desirable solution of performative prediction.

Performative stability. Performative prediction can be viewed as a game between the decision-maker and the population that responds to the deployed model. A first concept of optimality in this context is performative stability

$$\theta_{\mathrm{PS}} = \arg\min_{\theta} \mathrm{DPR}(\theta_{\mathrm{PS}}, \theta). \tag{2}$$

The expected loss over $\mathcal{D}(\theta_{\mathrm{PS}})$ is minimized at θ_{PS}. As such stable points constitute a fixed point of so-called *retraining* methods, which are heuristically adopted in dealing with distribution shifts. More specifically, at the t-th iteration, these methods find a set of parameters that minimizes the risk on the previously-generated distribution $\mathcal{D}(\theta_t)$. It can be formally described as $\theta_{t+1} = \arg\min_{\theta\in\Theta} \mathrm{DPR}(\theta_t, \theta)$. Several works have studied the convergence of population-based and stochastic retraining methods to stable points in the standard framework [9, 23, 29] as well as state-dependent extensions [4, 11, 17].

Performative optimality. An alternative solution concept is performative optimality. Performative optimal points describe models that achieve the minimal loss on the distribution they induce. In this sense, they are globally optimal in the following sense:

$$\theta_{\mathrm{PO}} = \arg\min_{\theta\in\Theta} \mathrm{DPR}(\theta, \theta).$$

It follows from the definition that $\mathrm{PR}(\theta_{\mathrm{PO}}) \leq \mathrm{PR}(\theta_{\mathrm{PS}})$ while in general, stable points are not necessarily optimal and optimal points are not necessarily stable. Furthermore, the two solutions θ_{PS} and θ_{PO} can be arbitrarily far apart from each other, see [25, 29]. Thus, retraining does not, in general, serve as a reliable heuristic for finding optimal points. That being said, we focus on finding performative optima in this work. There is only a handful of works that study approaches for finding optimal points. The first to tackle performative optimality was Miller et al. [25]. They proposed an iterative optimization method for minimizing the performative risk directly by first learning a model of the distribution map and then applying a gradient-based procedure to the resulting estimation of the performative risk. To guarantee global convergence of such an approach they studied structural assumptions under which the performative risk is convex. Around the same time, Izzo et al. [13] studied *performative gradient descent*, an algorithm that directly estimates the gradient of the performative risk. Technically, they assumed the existence of an estimator $\hat{f}(\theta)$ so that one can use samples from the distribution $\mathcal{D}(\theta)$ to infer θ, an assumption we will revisit later in this manuscript. To prove convergence to optimal points they directly assume convexity of the performative risk, an assumption that is hard to verify.

A different approach has recently been taken by Jagadeesan et al. [15] who leveraged tools from the bandits' literature for minimizing regret under performativity without making structural assumptions on the performative risk apart from weak regularity assumptions. Technically, they use data collected from experiments to build performative confidence bounds to characterize the regret

of unexplored parameters. Their method achieves sublinear regret and is the first to find optimal points for non-convex performative risk functions. However, a significant drawback of such a bandit approach is the need for global exploration, i.e., one needs to explore the entire parameter space to find the minimizer, something that is often not feasible in practical applications. Therefore, the focus of this work is on iterative descent methods that gradually approach optimal points.

2.3. Assumptions on the loss function

The loss function in performative prediction takes the same role as the loss in supervised learning. We discuss classical assumptions from the optimization literature such as strong convexity and smoothness on the loss function.

Assumption 1 (Smoothness) *The loss $\ell(\cdot;\cdot)$ is β-smooth if for all z, z', θ,*

$$\|\nabla_\theta \ell(z;\theta) - \nabla_\theta \ell(z';\theta)\|_2 \leq \beta \|z - z'\|_2.$$

Assumption 2 (Strong convexity) *The loss $\ell(\cdot;\cdot)$ is γ-strongly convex in θ if for all θ, θ', z,*

$$\ell(z;\theta') - \ell(z;\theta) \geq \nabla_\theta \ell(z;\theta)^\top (\theta' - \theta) + \frac{\gamma}{2}\|\theta - \theta'\|_2^2.$$

Note that this condition reduces to convexity when $\gamma = 0$.

These two assumptions characterize the dependence of the loss function on the model parameters θ and they are assumed to hold for any realization of z. We will adopt the smoothness assumption in this work and we will discuss the strong-convexity assumption as well as weaker notions of convexity.

In addition, we will need a regularity assumption for relating changes in the data distribution to changes in the loss function. Therefore, we will adopt the Lipschitzness condition w.r.t. z, which is common when dealing with performative distribution shifts [25, 29], namely

Assumption 3 (Lipschitzness) *The loss $\ell(\cdot;\cdot)$ is L-Lipschitz in z if for all z, z', θ,*

$$\|\nabla_z \ell(z;\theta) - \nabla_z \ell(z';\theta)\|_2 \leq L \|z - z'\|_2.$$

2.4. Characterizing the performative shift

Since performative effects are not under the control of the learner and typically unknown, we need assumptions on these distribution shifts for studying the problem and establishing convergence guarantees of algorithmic approaches. In contrast to classical supervised learning where optimization techniques are commonly adopted under conditions on the loss function, we need additional assumptions for performative optimization. This observation is at the core of the main difficulty in performative prediction and the art is to couple conditions on the loss function with conditions on the distribution shift to obtain favorable properties of the resulting performative risk minimization problem.

We shall first introduce the concept of Wasserstein distance, then we proceed to the key sensitivity assumption that relates changes in the model parameters to changes in the distribution.

Definition 1 (Wasserstein Distance) *The Wasserstein-1 distance, also called Earth-Mover (EM) distance, is defined as*

$$W(P_r, P_g) = \inf_{\gamma \in \Pi(P_r, P_g)} \mathbb{E}_{(x,y)\sim\gamma}\left[\, \|x - y\| \,\right], \tag{3}$$

where $\Pi(P_r, P_g)$ denotes the set of all joint distributions $\gamma(x, y)$ whose marginals are respectively P_r and P_g.

Intuitively, $\gamma(x, y)$ indicates how much mass must be transported from x to y in order to transform the distributions P_r into the distribution P_g. The EM distance then is the "cost" of the optimal transport plan. The infimum in Eq. 3 is highly intractable. On the other hand, the Kantorovich-Rubinstein duality tells us that

$$W(P_r, P_\theta) = \sup_{\|f\|_L \leq 1} \left[\int f(x) dP_r(x) - \int f(x) dP_\theta(x) \right]$$

This allows us to relate the distance of the expectation of a Lipschitz function over two different distributions by their Wasserstein distance.

Sensitivity. One commonly used assumption is a weak regularity condition that posits Lipschitzness of the distribution shift relative to changes in the model parameters. Intuitively, the idea is, if decisions are made according to similar predictive models, then the resulting distributions should also be close. More precisely, the *sensitivity condition* introduced by [29] relates changes in the model parameter as measured in Euclidean distance to changes in the distribution as measured in Wasserstein distance.

Assumption 4 (Sensitivity) *The distribution map $\mathcal{D}(\cdot)$ is ϵ-sensitive if for all θ, θ',*

$$W(\mathcal{D}(\theta), \mathcal{D}(\theta')) \leq \epsilon \|\theta - \theta'\|_2.$$

A prominent example of a distribution shift that satisfies the sensitivity assumption is strategic classification with quadratic costs [29]. Strategic classification microfounds performative prediction with the best response model for individual actions [14].

Mixture dominance. Below we present a stronger and more structural assumption on the distribution shift introduced by Miller et al. [25] which guarantees the convexity of $\mathrm{DPR}(\theta', \theta)$ w.r.t. the first argument θ'.

Assumption 5 (Mixture dominance) *A set of losses and functions satisfy the mixture dominance condition if $\mathrm{DPR}(\theta', \theta)$ is convex w.r.t. θ', thus for all θ, θ',*

$$\mathbb{E}_{z\sim\mathcal{D}(\theta)}[\ell(z;\theta)\nabla_\theta \log\left(p_\theta(z)\right)]^\top(\theta' - \theta) \leq \mathrm{DPR}(\theta', \theta) - \mathrm{PR}(\theta).$$

It has been demonstrated by [25] that this assumption is satisfied in performative environments by the well-known *location-scale* distribution family, which is a family of distributions formed by translation and re-scaling of a standard family member. For example, let the underlying distribution be $\mathcal{D}_\theta$, data points $z_\theta \sim \mathcal{D}_\theta$ satisfies $z_\theta = z_0 + A\theta$, where the matrix A is an unknown parameter, and z_0 is is a zero-mean random variable. Here the additive term $A\theta$ represents the effects of performativity. Standard examples include Gaussian, Cauchy and uniform distributions. There have been wide applications for these families in economic [35] and statistics [12]. Examples of location families can also be found in strategic classification [10].

3. Beyond convexity of the performative risk

In studying the convergence of gradient-based methods to performative optima, previous works on performative prediction have focused on a setting where the performative risk is convex. While some works such as [13] directly start from apriori hard-to-verify assumptions, the notable work by Miller et al. [25] provides conditions on the loss function and the distribution map under which global convexity of PR emerges. More specifically they show that strong convexity of the loss function, together with Assumptions 4-5 to control the performativity effects of distributions, is sufficient to establish (strong) convexity of the performative risk. They obtain a zeroth-order algorithm that provably converges towards globally performative optimal points (θ_{PO}) [25], which could avoid generally difficult computing performative gradients.

In the following, we take a step further by demonstrating that the strong convexity assumption on the loss function can be relaxed while preserving the amenability of the performative risk minimization problem to gradient-based optimization. We build on the existing literature on first-order optimization that showed that convexity can be relaxed to weaker conditions while maintaining the same convergence guarantees of gradient-based optimizers.

3.1. Weaker notions of strong convexity

In the optimization literature, strongly-convex optimization objectives are playing an important role as they can be optimized using gradient-descent methods at a linear rate. In the context of finding performatively stable points, this linear rate can be maintained by retraining methods in the presence of weak performative shifts [29].

There are various works in the optimization literature that aim to relax the strong-convexity condition on the objective function while maintaining favorable convergence properties. Examples include error bounds (EB) [8, 21], essential strong convexity (ESC) [19, 32], weak strong-convexity (WSC) [22, 26], restricted secant inequality (RSI) [37, 38], restricted strong-convexity (RSC) [27, 38], Polyak-Lojasiewicz (PL) inequality [20, 30] and quadratic growth (QG) condition [1, 8]. The relations between these conditions have been studied in [3, 16, 18, 37]. In particular, for a smooth function with Lipschitz-continuous gradients, Karimi et al. [16] showed the following chain of implications:

$$(SC) \to (ESC) \to (WSC) \to (RSI) \to (EB) \equiv (PL) \to (QG). \tag{4}$$

This implication chain is widely used in numerical optimization [28, 33, 34].

We build on this work to analyze the performative prediction from an optimization perspective. Previous works assumed the loss ℓ satisfies strong convexity (SC) globally which is the strongest condition according to the implication chain in Eq. 4. Our contribution is to show that weaker conditions on DPR still imply desirable local properties of PR. Specifically, in this work, we prove that assuming WSC and RSI of DPR is sufficient to establish WC and RSI of PR, respectively, following similar steps taken in Miller et al. [25]. These local conditions only need to hold around the minimizer and do not posit strong structural restrictions on the overall landscape such as SC, thus making the assumptions more realistic in practice.

3.2. Establishing weak convexity of the performative risk

First, we introduce the *weak strong convexity* [16, 26] in order to relax the strong convexity conditions. We will follow the convention that x_p denotes the projection of x onto the optimal solution set $\mathcal{X}^*$ throughout this paper.

Definition 2 (WSC) *A function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-WSC if for all x we have*

$$f^* \geq f(x) + \langle \nabla f(x), x_p - x \rangle + \frac{\mu}{2}\|x_p - x\|^2. \tag{5}$$

where $x_p \in \mathcal{X}^* = \arg\min_x f(x)$ and $f^* = f(x_p)$.

We translate this definition to the decoupled performative risk as follows:

Assumption 6 *For performative optimum θ_{PO} and its induced distribution $\mathcal{D} = \mathcal{D}(\theta_{\mathrm{PO}})$, suppose the optimal solution for minimizing $\mathrm{DPR}(\theta_{\mathrm{PO}}, \cdot)$ is θ^*. We say $\mathrm{DPR}(\theta_{\mathrm{PO}}, \cdot)$ satisfies $\mu-$WSC, if for any $\theta \in \Theta$ it holds that*

$$\mathrm{DPR}(\theta_{\mathrm{PO}}, \theta^*) \geq \mathrm{DPR}(\theta_{\mathrm{PO}}, \theta) + \nabla_\theta \mathrm{DPR}(\theta_{\mathrm{PO}}, \theta)^\top (\theta^* - \theta) + \frac{\mu}{2}\|\theta^* - \theta\|^2. \tag{6}$$

This definition requires that the minimizer for the expected loss on the distribution $\mathcal{D}$ is unique: $\theta^* = \arg\min_{\theta \in \Theta} \mathrm{DPR}(\theta_{\mathrm{PO}}, \theta)$. This is a mild assumption used by previous work focusing on repeated optimization in performative environments [9, 23, 36]. Intuitively, we posit a local property on the structure of DPR over the *specific* distribution $\mathcal{D}$ induced by θ_{PO}. The quantity is measured over the domain of model parameter θ. We note that this condition is already weaker than the strong convexity Assumption 2 which is assumed to be true for *any data point* z [25], thus depending heavily on the structure of the loss function.

Equipped with Assumption 6, we are ready to state the following new theorem.

Theorem 3 *Under Assumptions 1, 4-6, suppose the performative optimum is also performative stable, then the performative risk is guaranteed to be weakly convex if $\frac{\mu}{2\beta} \geq \epsilon$.*

We make justification on the assumption that θ_{PO} is also performative stable. As mentioned, optimal points and stable points do not generally imply each other. Therefore, simply adopting retraining (e.g., [29]) does not return global optima. Nevertheless, the local properties of $\mathrm{PR}(\theta)$ under this assumption would be really helpful in determining the optimality of a performative stable point. In the next section, we will extend to more results on what information about θ_{PO} we could obtain at a specific stable point without the assumption.

Proof [Proof sketch] Technically, our derivation starts from Eq. 6. After expanding the inequality around the minimizer of θ_{PO}, we split $\nabla\mathrm{PR}$ into two terms: gradient of the loss function $\mathbb{E}_{z\sim\mathcal{D}(\theta)}\nabla_\theta \ell(z;\theta)$, and gradient on the variable distribution $\mathbb{E}_{z\sim\mathcal{D}(\theta)}\ell(z;\theta)\nabla_\theta \log p_\theta(z)$, where p_θ is the probability distribution function of $\mathcal{D}(\theta)$. We then use the sensitivity Assumption 4 to bound the first term by the quadratic distance $\|\theta_{\mathrm{PO}} - \theta\|^2$ and use the mixture dominance Assumption 5 to control the second term. ■

The $\frac{\mu}{2\beta}$ term in the theorem is a sharp threshold for weak convexity of the performative risk, which aligns with the observations in [25, 29]. See details in Appendix A.

3.3. Establishing RSI of the performative risk

The second condition we consider is that *Restricted Secant Inequality* [38]. It is defined as follows:

Definition 4 (RSI) *A function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies Restricted Secant Inequality (RSI) if for all x we have*

$$\langle \nabla f(x), x - x_p \rangle \geq \mu \|x_p - x\|^2. \tag{7}$$

Again, x_p is the projection of x onto the optimal solution set $\mathcal{X}^*$.

Remark 5 *A function satisfies restricted strongly-convex (RSC) if it is convex and also satisfies RSI.*

Assumption 7 *For performative optimum θ_{PO} and its induced distribution $\mathcal{D}$, suppose the optimal solution for minimizing $\mathrm{DPR}(\theta_{\mathrm{PO}}, \cdot)$ is unique and is denoted as θ^*. We say $\mathrm{DPR}(\theta_{\mathrm{PO}}, \cdot)$ satisfies $\mu-$RSI, if for any θ it holds*

$$\langle \nabla_\theta \mathrm{DPR}(\theta_{\mathrm{PO}}, \theta), \theta - \theta^* \rangle \geq \mu \|\theta^* - \theta\|^2 \tag{8}$$

Theorem 6 *Under Assumptions 3-5, and 7, suppose the performative optima is also performative stable, then $\mathrm{PR}(\theta)$ satisfies RSI.*

Note that compared to WSC, the RSI condition is weaker and thus generalizes more easily in first-order optimization. See details in Appendix A. We conjecture that specific implementations discussed by Negahban and Wainwright [27], Zhang [37], Zhang and Yin [38] could be extended to optimize the PR in this setting.

4. When are local properties sufficient for stability and optimality?

As aforementioned, existing methods in the literature that seek stable points do not have global optimality guarantees. Meanwhile, current methods aiming at optimizing PR *directly* require local structural assumptions around the performative optima that may not hold in real settings.

In this section, we investigate relations between stable points and optimal points. We relate to previous work focusing on the convergence of retraining methods to stable points [2, 9, 23, 29, 36] and argue that if stable points offer a good approximation for optimal points, then they can serve as a good starting point to ensure convergence to optimal points with only local assumption on the performative risk.

Suppose the Lipschitzness condition for loss (c.f. Assumption 3) holds throughout this section, which is standard and mild in learning theory or optimization theory literature. For any θ, θ' in the parameter domain, we could quantify the gap between $\mathrm{PR}(\theta)$ and $\mathrm{PR}(\theta')$ as

$$\begin{aligned}
\mathrm{PR}(\theta') &= \mathrm{DPR}(\theta', \theta') \\
&\geq \mathrm{DPR}(\theta, \theta') - LW(\mathcal{D}(\theta), \mathcal{D}(\theta')) \\
&= \mathrm{PR}(\theta) + [\mathrm{DPR}(\theta, \theta') - \mathrm{DPR}(\theta, \theta)] - LW(\mathcal{D}(\theta), \mathcal{D}(\theta')).
\end{aligned}$$

The second term on the RHS measures the distance in the function value between θ and θ' over the distribution $\mathcal{D}(\theta)$.

In the remainder of this section, we pick θ to be a stable point (i.e., $\theta = \arg\min_x \mathrm{DPR}(\theta, x)$) so we can work with the suboptimality gap of the function. First rewrite the above equation as

$$\mathrm{PR}(\theta') \geq \mathrm{PR}(\theta) + \Delta_\theta(\theta') - LW(\mathcal{D}(\theta), \mathcal{D}(\theta')). \tag{9}$$

where $\Delta_\theta(\theta') := \mathrm{DPR}(\theta, \theta') - \mathrm{PR}(\theta) \geq 0$ is the suboptimality of θ' measured over $\mathcal{D}(\theta)$

Proposition 7 *Under Assumption 3, if*

$$\Delta_\theta(\theta') \geq LW(\mathcal{D}(\theta), \mathcal{D}(\theta')), \tag{10}$$

it holds that

$$\mathrm{PR}(\theta) \leq \mathrm{PR}(\theta').$$

The proof is straightforward by noticing that, when Eq. 10 holds, we have $\mathrm{PR}(\theta) \leq \mathrm{PR}(\theta')$ for all $\theta' \in \Theta$. Thus θ is performative optimal according to definition. Intuitively, Eq. 10 relates changes in the loss to changes in the distribution.

Starting from θ. Suppose we have reached θ which is a performative stable point. To achieve θ_{PO}, we may consider two directions. (1) How is the performance of $\mathrm{PR}(\theta)$? What is the gap between $\mathrm{PR}(\theta)$ and the optimal performative risk (denoted as PR^*)? (2) What geometric information about θ_{PO} could we obtain at θ? As we may believe that θ will serve as a good approximation for θ_{PO} when the Euclidean distance is close. We believe investigating along the two directions is promising since it makes use of the large amount of work focusing on reaching θ_{PS}.

Below, we give three preliminary examples on how some general structural conditions on the loss function and distribution maps could lead to useful results related to the two directions.

We show the suboptimality of θ under Lipschitzness and a type of boundedness.

Example 1 *Assume performative shifts are bounded by an absolute value, i.e.,*

$$W(\mathcal{D}(\theta), \mathcal{D}(\theta')) \leq B.$$

We have the following bound characterizing the suboptimality of θ

$$\mathrm{PR}(\theta) - \mathrm{PR}^* \leq LB. \tag{11}$$

Proof The proof is straightforward by noticing $\Delta_\theta(\theta') \geq 0$ when θ being stable. From Eq. 9 we have $\mathrm{PR}(\theta') \geq \mathrm{PR}(\theta) - LB$. ■

The next two examples will show that θ_{PO} could not be far from θ.

Example 2 *Assume (1) performative shifts are bounded by an absolute value B and (2) $\Delta_\theta(\theta')$ satisfies quadratic growth, i.e., $\Delta_\theta(\theta') \geq \gamma\|\theta - \theta'\|^2$, we have that performative optimal point θ_{PO} satisfies*

$$\|\theta_{\mathrm{PO}} - \theta\| \leq \sqrt{\frac{LB}{\gamma}}.$$

Proof We proof by contradiction. Suppose there exists a performative optimal point θ_{PO} which is $\sqrt{LB/\gamma}-$close to θ. The quadratic growth shows

$$\Delta_\theta(\theta_{\mathrm{PO}}) \geq LB \geq LW(\mathcal{D}(\theta), \mathcal{D}(\theta_{\mathrm{PO}})).$$

From Eq. 9 we have $\mathrm{PR}(\theta_{\mathrm{PO}}) \geq \mathrm{PR}(\theta)$ which contradicts with the optimality of θ_{PO}. ■

Example 3 *Under Assumption 4, suppose $\Delta_\theta(\theta')$ satisfies quadratic growth, the performative optimal point θ_{PO} satisfies*

$$\|\theta_{\mathrm{PO}} - \theta\| \leq \frac{L\epsilon}{\gamma}.$$

Proof We proof by contradiction. Suppose there exists a performative optimal point θ_{PO} which is $L\epsilon/\gamma-$close to θ. The quadratic growth shows $\Delta_\theta(\theta_{\mathrm{PO}}) \geq \frac{L^2\epsilon^2}{\gamma}$. At the meantime, senstitivity assumption 4 shows

$$W(\mathcal{D}(\theta), \mathcal{D}(\theta_{\mathrm{PO}})) \leq \epsilon\|\theta - \theta_{\mathrm{PO}}\| \leq \frac{L\epsilon^2}{\gamma}.$$

Therefore we have $\Delta_\theta(\theta_{\mathrm{PO}}) \geq LW(\mathcal{D}(\theta), \mathcal{D}(\theta_{\mathrm{PO}}))$. Substituting the relation into Eq. 9 shows $\mathrm{PR}(\theta_{\mathrm{PO}}) \geq \mathrm{PR}(\theta)$ which contradicts with the optimality of θ_{PO}. ■

To conclude, in this section we attempt to make use of stable points in finding optima points. Though θ_{PS} and θ_{PO} do not imply each other as discussed aforementioned, we aim at revealing when and how stable points could be served as starting points for finding optima points. In such scenario, previous work focusing on convergence of retraining methods to stable points would still be worthy.

Concretely, we have shown several conditions under which stable points offer a good approximation for optimal points. Then, with only local assumption on PR, we show that one could start from stable points in an optimization procedure, and finally, converge to optimal points.

Remark 8 *An interesting special case where stable points are optimal without requiring some sort of global optimality is when the Bayesian error of every distribution is the same. This means all minima are on one level-set. Hence, no other point can have a smaller loss.*

5. Conclusion

This paper studies optimization aspects of the performative risk minimization problem.

First, we relax the standard strong convexity assumption on the loss function required in prior work to show convergence of iterative optimization methods in performative prediction. In particular, we study two weaker conditions on DPR: WSC and RSI – both are sufficient to establish weak convexity of the performative risk under suitable assumptions on the distribution map. Our work takes a first step towards importing advanced assumptions from the classical optimization literature into performative prediction.

Second, we make a contribution by raising interesting questions about the meaning of local regularity assumptions in the context of performative prediction. If stable points can be found heuristically and serve as a good approximation for global optima, the local regularity of PR can be sufficient to find optimal points. We provide some bounds on the distance between optimal and stable points but it remains to better understand how they relate in practical settings.

As of future work, we note that a very interesting and important topic to study is the PL condition which is among the weakest assumptions according to the implication chain and is ubiquitous in over-parameterized neural networks [39].

Suppose the following μ-PL condition holds

$$\frac{1}{2}\left\|\nabla_{\theta'}\mathrm{DPR}(\theta, \theta')\right\|^2 \geq \mu(\mathrm{DPR}(\theta, \theta') - \mathrm{PR}(\theta)) = \mu\Delta_\theta(\theta').$$

Naturally, we expand the RHS through Eq. 9

$$\frac{\|\nabla_{\theta'}\mathrm{DPR}(\theta,\theta')\|^2}{2\mu} \geq \mathrm{PR}(\theta') - \mathrm{PR}(\theta) - LW(\mathcal{D}(\theta),\mathcal{D}(\theta'));.$$

Intuitively, we raise several interesting questions related to finding performative optima.

1. Understanding when and how (e.g., some structural properties of loss function or a natural set of distributions), it holds that $W(\mathcal{D}(\theta),\mathcal{D}(\theta')) \leq C\|\nabla_{\theta'}\mathrm{DPR}(\theta,\theta')\|^2$ where C is a certain constant. Since the condition characterizes local properties of DPR near performative stable points, it could be more common in practice. Notice that current performative prediction literature heavily rely on sensitivity relation (c.f. Assumption 4).

2. What is the impact of data pre-processing steps on the implications of performative shifts? Can they influence performativity effects? For example, can data whitening, result in distribution with favorable properties, such as location-scale family?

References

[1] Mihai Anitescu. Degenerate nonlinear programming with a quadratic growth condition. *SIAM Journal on Optimization*, 10(4):1116–1135, 2000.

[2] Gianluca Bianchin, Miguel Vaquero, Jorge Cortes, and Emiliano Dall'Anese. Online stochastic optimization for unknown linear systems: Data-driven synthesis and controller analysis. *arXiv preprint arXiv:2108.13040*, 2021.

[3] Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

[4] Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative Prediction in a Stateful World. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 6045–6061. PMLR, 2022.

[5] Michel Callon et al. What does it mean to say that economics is performative. *Do economists make markets*, pages 311–357, 2007.

[6] Chris Clarke. Financial engineering, not economic photography: Popular discourses of finance and the layered performances of the sub-prime crisis. *Journal of Cultural Economy*, 5(3): 261–278, 2012.

[7] Franck Cochoy, Martin Giraudeau, and Liz McFall. Performativity, economics and politics: An overview. *Journal of Cultural Economy*, 3(2):139–146, 2010.

[8] Dmitriy Drusvyatskiy and Adrian S Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

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

[10] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

[11] Keegan Harris, Hoda Heidari, and Steven Z Wu. Stateful strategic regression. *Advances in Neural Information Processing Systems*, 34:28728–28741, 2021.

[12] Nil Kamal Hazra, Mithu Rani Kuiti, Maxim Finkelstein, and Asok K Nanda. On stochastic comparisons of maximum order statistics from the location-scale family of distributions. *Journal of Multivariate Analysis*, 160:31–41, 2017.

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

[14] Meena Jagadeesan, Celestine Mendler-Dünner, and Moritz Hardt. Alternative microfoundations for strategic classification. In *International Conference on Machine Learning*, pages 4687–4697. PMLR, 2021.

[15] Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. Regret Minimization with Performative Feedback. In *International Conference on Machine Learning*, pages 9760–9785. PMLR, 2022.

[16] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

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

[18] Ji Liu and Stephen J Wright. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *SIAM Journal on Optimization*, 25(1):351–376, 2015.

[19] Ji Liu, Steve Wright, Christopher Ré, Victor Bittorf, and Srikrishna Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. In *International Conference on Machine Learning*, pages 469–477. PMLR, 2014.

[20] Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 1306–1314. PMLR, 2021.

[21] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

[22] Chenxin Ma, Rachael Tappenden, and Martin Takáč. Linear convergence of randomized feasible descent methods under the weak strong convexity assumption. *The Journal of Machine Learning Research*, 17(1):8138–8161, 2016.

[23] Celestine Mendler-Dünner, Juan C. Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic Optimization for Performative Prediction. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, 2020.

[24] Celestine Mendler-Dünner, Frances Ding, and Yixin Wang. Predicting from Predictions. *arXiv preprint arXiv:2208.07331*, 2022.

[25] John P Miller, Juan C Perdomo, and Tijana Zrnic. Outside the Echo Chamber: Optimizing the Performative Risk. In *International Conference on Machine Learning*, pages 7710–7720. PMLR, 2021.

[26] Ion Necoara, Yu Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1):69–107, 2019.

[27] Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1): 1665–1697, 2012.

[28] Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[29] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

[30] Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki*, 3(4):643–653, 1963.

[31] Brian Roberts. Performative social science: A consideration of skills, purpose and context. *Historical Social Research/Historische Sozialforschung*, pages 307–353, 2009.

[32] Thomas Strömberg. Duality between Fréchet differentiability and strong convexity. *Positivity*, 15(3):527–536, 2011.

[33] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204. PMLR, 2019.

[34] Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30, 2017.

[35] Wing-Keung Wong and Chenghu Ma. Preferences over location-scale family. *Economic Theory*, 37(1):119–146, 2008.

[36] Killian Wood, Gianluca Bianchin, and Emiliano Dall'Anese. Online projected gradient descent for stochastic optimization with decision-dependent distributions. *IEEE Control Systems Letters*, 6:1646–1651, 2021.

[37] Hui Zhang. The restricted strong convexity revisited: analysis of equivalence to error bound and quadratic growth. *Optimization Letters*, 11(4):817–833, 2017.

[38] Hui Zhang and Wotao Yin. Gradient methods for convex minimization: better rates under weaker conditions. *arXiv preprint arXiv:1303.4645*, 2013.

[39] Mo Zhou, Rong Ge, and Chi Jin. A local convergence theory for mildly over-parameterized two-layer neural network. In *Conference on Learning Theory*, pages 4577–4632. PMLR, 2021.

Appendix A. Proofs for Section 3

[Proofs for Theorem 3]

Proof At θ_{PO}, since performative optimum is also performative stable, it is known $\theta^* = \theta_{\mathrm{PO}}$.

Using WSC property we have

$$\mathbb{E}_{z\sim\mathcal{D}(\theta_{\mathrm{PO}})}\ell(z;\theta_{\mathrm{PO}}) \geq \mathbb{E}_{z\sim\mathcal{D}(\theta_{\mathrm{PO}})}\ell(z;\theta) + \langle \nabla_\theta \mathrm{DPR}(\theta_{\mathrm{PO}},\theta), \theta_{\mathrm{PO}} - \theta\rangle + \frac{\mu}{2}\|\theta_{\mathrm{PO}} - \theta\|^2. \quad (12)$$

Our goal is to characterize the function class of $\mathrm{PR}(\theta) = \mathbb{E}_{z\sim\mathcal{D}(\theta)}[l(z;\theta)]$

Potentially, we expect it to be weak convex w.r.t. θ, i.e.

$$\mathrm{PR}(\theta_{\mathrm{PO}}) \geq \mathrm{PR}(\theta) + \langle \nabla \mathrm{PR}(\theta), \theta_{\mathrm{PO}} - \theta\rangle, \quad (13)$$

or using the definition of $\mathrm{PR}(\theta)$:

$$\mathbb{E}_{z\sim\mathcal{D}(\theta_{\mathrm{PO}})}[l(z;\theta_{\mathrm{PO}})] \geq \mathbb{E}_{z\sim\mathcal{D}(\theta)}[l(z;\theta)] + \langle \nabla \mathbb{E}_{z\sim\mathcal{D}(\theta)}[l(z;\theta)], \theta_{\mathrm{PO}} - \theta\rangle. \quad (14)$$

First, we separate the gradient of PR into

$$\nabla_\theta \mathrm{PR}(\theta) = \overbrace{\mathbb{E}_{z\sim\mathcal{D}(\theta)}\nabla_\theta \ell(z;\theta)}^{\nabla_1} + \overbrace{\mathbb{E}_{z\sim\mathcal{D}(\theta)}\ell(z;\theta)\nabla_\theta \log p_\theta(z)}^{\nabla_2}. \quad (15)$$

We have

$$\begin{aligned}
&\mathrm{PR}(\theta_{\mathrm{PO}}) \geq \mathrm{PR}(\theta) + \langle \nabla \mathrm{PR}(\theta), \theta_{\mathrm{PO}} - \theta\rangle \\
&\overset{(i)}{\Leftarrow} \mathrm{DPR}(\theta_{\mathrm{PO}},\theta) \geq \mathrm{PR}(\theta) + \nabla^\top(\theta_{\mathrm{PO}} - \theta) - \langle \mathbb{E}_{z\sim\mathcal{D}(\theta_{\mathrm{PO}})}\nabla\ell(z;\theta), \theta_{\mathrm{PO}} - \theta\rangle - \frac{\mu}{2}\|\theta_{\mathrm{PO}} - \theta\|^2 \\
&\Leftrightarrow \mathrm{DPR}(\theta_{\mathrm{PO}},\theta) \geq \mathrm{PR}(\theta) + \nabla_2^\top(\theta_{\mathrm{PO}} - \theta) + \langle \nabla_1 - \mathbb{E}_{z\sim\mathcal{D}(\theta_{\mathrm{PO}})}\nabla\ell(z;\theta), \theta_{\mathrm{PO}} - \theta\rangle - \frac{\mu}{2}\|\theta_{\mathrm{PO}} - \theta\|^2 \\
&\overset{(ii)}{\Leftarrow} \mathrm{DPR}(\theta_{\mathrm{PO}},\theta) \geq \mathrm{PR}(\theta) + \nabla_2^\top(\theta_{\mathrm{PO}} - \theta) + (\beta\epsilon - \frac{\mu}{2})\|\theta_{\mathrm{PO}} - \theta\|^2 \\
&\overset{(iii)}{\Leftarrow} \frac{\mu}{2} \geq \beta\epsilon
\end{aligned}$$

(i). Here we use: $\mathrm{PR}(\theta_{\mathrm{PO}}) \geq \mathrm{DPR}(\theta_{\mathrm{PO}},\theta) + \langle \mathbb{E}_{z\sim\mathcal{D}(\theta_{\mathrm{PO}})}\nabla\ell(z;\theta), \theta_{\mathrm{PO}} - \theta\rangle + \frac{\mu}{2}\|\theta_{\mathrm{PO}} - \theta\|^2$.

(ii). This is because ϵ-sensitive of distribution map $\mathcal{D}$.

(iii). Mixture dominance implies: $\mathrm{DPR}(\theta_{\mathrm{PO}},\theta) \geq \mathrm{PR}(\theta) + \nabla_2^\top(\theta_{\mathrm{PO}} - \theta)$.

We conclude that we only need weak strong convexity of DPR, local mixture dominance and local distribution sensitivity to guarantee that PR is weakly-convex to θ_{PO}, instead of strongly convex everywhere. ■

[Proofs for Theorem 6]

Proof At θ_{PO}, because performative optimum is also performative stable, it is known $\theta^* = \theta_{\mathrm{PO}}$.

Using RSI property we have

$$\langle \nabla_\theta \mathrm{DPR}(\theta_{\mathrm{PO}},\theta), \theta - \theta_{\mathrm{PO}}\rangle \geq \mu\|\theta_{\mathrm{PO}} - \theta\|^2.$$

We derive back as following, we want to show PR is RSI, i.e.,

$$
\begin{aligned}
&\langle \nabla_\theta \mathrm{PR}(\theta), \theta - \theta_{\mathrm{PO}} \rangle \geq \mu' \|\theta_{\mathrm{PO}} - \theta\|^2 \\
\Leftrightarrow\ & 0 \geq \mu' \|\theta_{\mathrm{PO}} - \theta\|^2 + \langle \nabla_1, \theta_{\mathrm{PO}} - \theta \rangle + \langle \nabla_2, \theta_{\mathrm{PO}} - \theta \rangle \\
\Leftrightarrow\ & 0 \geq \mu' \|\theta_{\mathrm{PO}} - \theta\|^2 + \langle \nabla_1 - \nabla_\theta \mathrm{DPR}(\theta_{\mathrm{PO}}, \theta), \theta_{\mathrm{PO}} - \theta \rangle + \langle \nabla_\theta \mathrm{DPR}(\theta_{\mathrm{PO}}, \theta), \theta_{\mathrm{PO}} - \theta \rangle + \langle \nabla_2, \theta_{\mathrm{PO}} - \theta \rangle \\
\overset{(i)}{\Leftarrow}\ & 0 \geq (\mu' + \beta\epsilon) \|\theta_{\mathrm{PO}} - \theta\|^2 + \langle \nabla_\theta \mathrm{DPR}(\theta_{\mathrm{PO}}, \theta), \theta_{\mathrm{PO}} - \theta \rangle + \langle \nabla_2, \theta_{\mathrm{PO}} - \theta \rangle \\
\overset{(ii)}{\Leftarrow}\ & 0 \geq (\mu' + \beta\epsilon) \|\theta_{\mathrm{PO}} - \theta\|^2 + \langle \nabla \mathrm{DPR}(\theta_{\mathrm{PO}}, \theta), \theta_{\mathrm{PO}} - \theta \rangle + \mathrm{DPR}(\theta_{\mathrm{PO}}, \theta) - \mathrm{PR}(\theta) \\
\overset{(iii)}{\Leftarrow}\ & 0 \geq (\mu' + \beta\epsilon + L\epsilon) \max(\|\theta_{\mathrm{PO}} - \theta\|, \|\theta_{\mathrm{PO}} - \theta\|^2) + \langle \nabla \mathrm{DPR}(\theta_{\mathrm{PO}}, \theta), \theta_{\mathrm{PO}} - \theta \rangle
\end{aligned}
$$

because

(i). Because we have ϵ-sensitiveness of distribution map $\mathcal{D}$.

(ii). From mixture dominance we know $\mathrm{DPR}(\theta_{\mathrm{PO}}, \theta) \geq \mathrm{PR}(\theta) + \langle \nabla_2, \theta_{\mathrm{PO}} - \theta \rangle$.

(iii). This is because of L-Lipschitzness of loss function w.r.t. z.

We have 1) If $\|\theta_{\mathrm{PO}} - \theta\|^2 \leq \|\theta_{\mathrm{PO}} - \theta\|$ (close to optimum), then

$$
\underbrace{\langle \nabla \mathrm{DPR}(\theta_{\mathrm{PO}}, \theta), \theta - \theta_{\mathrm{PO}} \rangle}_{\geq \mu \|\theta_{\mathrm{PO}} - \theta\|^2} \geq (\mu' + \beta\epsilon + L\epsilon) \|\theta_{\mathrm{PO}} - \theta\| \tag{16}
$$

The retrodiction shows when assuming DPR is μ-RSI, then our goal is achieved if

$$
\mu' = \mu \|\theta_{\mathrm{PO}} - \theta\| - (\beta + L)\epsilon \geq 0 \tag{17}
$$

2) If $\|\theta_{\mathrm{PO}} - \theta\| \leq \|\theta_{\mathrm{PO}} - \theta\|^2$ then

$$
\langle \nabla \mathrm{DPR}(\theta_{\mathrm{PO}}, \theta), \theta - \theta_{\mathrm{PO}} \rangle \geq \mu \|\theta_{\mathrm{PO}} - \theta\|^2 \geq (\mu' + \beta\epsilon + L\epsilon) \|\theta_{\mathrm{PO}} - \theta\|^2 \tag{18}
$$

The retrodiction shows: when assuming DPR is μ-RSI, then our goal is achieved if

$$
\mu' = \mu - (\beta + L)\epsilon \geq 0 \tag{19}
$$

Proof is completed. ■