
Statistical Inference under Performativity

Xiang Li*

Independent Researcher
xiangli0233@gmail.com

Yunai Li*

Northwestern University
yunaili2030@u.northwestern.edu

Huiying Zhong*

MIT
zhong826@mit.edu

Lihua Lei

Stanford University
lihuallei@stanford.edu

Zhun Deng

UNC at Chapel Hill
zhundeng@cs.unc.edu

Abstract

Performativity of predictions refers to the phenomenon where prediction-informed decisions influence the very targets they aim to predict—a dynamic commonly observed in policy-making, social sciences, and economics. In this paper, we *initiate* an end-to-end framework of statistical inference under performativity. Our contributions are twofold. First, we establish a central limit theorem for estimation and inference in the performative setting, enabling standard inferential tasks such as constructing confidence intervals and conducting hypothesis tests in policy-making contexts. Second, we leverage this central limit theorem to study prediction-powered inference (PPI) under performativity. This approach yields more precise estimates and tighter confidence regions for the model parameters (i.e., policies) of interest in performative prediction. We validate the effectiveness of our framework through numerical experiments. To the best of our knowledge, this is the first work to establish a complete statistical inference under performativity, introducing new challenges and inference settings that we believe will provide substantial value to policy-making, statistics, and machine learning.

1 Introduction

Prediction-informed decisions are ubiquitous in nearly all areas and play important roles in our daily lives. An important and commonly observed phenomenon is that prediction-informed decisions can impact the targets they aim to predict, which is called *performativity* of predictions. For instance, policies about loans based on default risk prediction can alter consumption habits of the population that will further have an impact on their ability to pay off their loans.

To characterize performativity of predictions, a rich line of work on performative prediction [25] have been formalizing and investigating this idea that predictive models used to support decisions can impact the data-generating process. Mathematically, given a parameterized loss function ℓ , the aim of performative prediction is to optimize the performative risk:

$$\text{PR}(\theta) := \mathbb{E}_{z \sim \mathcal{D}(\theta)} \ell(z; \theta) \quad (1)$$

where $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$ is the input and output pair drawn from a distribution $\mathcal{D}(\theta)$ that is dependent on the loss parameter θ . Typically, $\mathcal{D}(\theta)$ is unknown and the optimization objective $\text{PR}(\theta)$ can be non-convex even if $\ell(z; \theta)$ is convex in θ . Thus, finding a *performative optimal* point $\theta_{\text{PO}} \in \arg \min_{\theta} \text{PR}(\theta)$ (there might exist multiple optimizers due to non-convexity) can be theoretically intractable unless we impose very strong distributional assumptions [20]. As an

* for equal contribution, listed in alphabetical order.

alternative, [25] mainly study how to obtain a *performative stable* point, which satisfies the following relationship:

$$\theta_{\text{PS}} = \arg \min_{\theta} \mathbb{E}_{z \sim \mathcal{D}(\theta_{\text{PS}})} \ell(z; \theta).$$

The performative stable point could be proved unique under some regularity conditions, and it could be shown close to a performative optimal point when the distribution shift between different θ 's is not too dramatic, which makes it a good proxy to θ_{PO} . In particular, θ_{PS} could be considered as a good proxy to the Stackelberg equilibrium in the strategic classification setting. Moreover, it could be calculated in distribution-agnostic settings.

Previous work mainly focuses on prediction performance and convergence rate analysis for performative prediction. On the contrary, another important aspect, ***inference under performativity***, eludes the literature. Although a central limit theorem was established in [7] for the stochastic gradient update algorithm with a single sample under performativity (a setting that is often impractical in applications such as policy-making), the authors assumed all structural knowledge to be given and did not provide any data-driven methods for estimating the covariance of the various quantities appearing in their central limit theorem. Therefore, [7] ***did not provide a complete statistical inference framework under performativity***.

However, inference is extremely important in performative prediction because parameter θ in many scenarios represents a concrete policy, such as a tax rate or credit score cutoff. Thus, when it comes to policy-making, the aim of tackling $\text{PR}(\theta)$ is not just for prediction, but more for obtaining a good policy. As a result, knowing convergence to θ_{PS} is not enough, we need to build statistical inference for θ_{PS} so as to enable people to report additional critical information like confidence or conduct necessary hypothesis testing.

Our contributions. In light of the importance of building statistical inference under performativity, in this work, we build a framework including the following elements.

- (1). As our first contribution, we investigate a widely applied iterative algorithm to calculate θ_{PS} , i.e., repeated risk minimization (RRM) (see details in Section 2), and establish central limit theorems for the $\hat{\theta}_t$'s obtained in the RRM process towards θ_{PS} . Based on that, we are able to obtain the confidence region for θ_{PS} . Our results could be viewed as generalizing standard statistical inference from a *static* setting to a *dynamic* setting.
- (2). As our second contribution, we further leverage the derived central limit theorems to investigate prediction-powered inference (PPI), another recently popular topic in modern statistical learning, under performativity. Our results generalize previous work [1] to a dynamic performative setting. This enables us to obtain better estimation and inference for the RRM process and θ_{PS} . More importantly, our results could also help mitigate ***data scarcity*** issues in getting feedback about policy implementation that often conducted by doing surveys that frequently encounter non-responses [13]. Thus, we also contribute to generalizing the line of work on performative prediction by introducing a ***more data efficient*** algorithm.

To sum up, our work establishes the first end-to-end framework for statistical inference under performativity for the celebrated repeated risk minimization algorithm. Meanwhile, we introduce prediction-powered inference under performativity to enable a more efficient inference. We believe our work would inspire new interesting topics and bring up new challenges to both areas of performative prediction and prediction-powered inference, as well as add significant value to policy-making in a broad range of areas such as social science and economics.

1.1 Related Work

Performative prediction. Performativity describes the phenomenon whereby predictions influence the outcomes they aim to predict. [25] were the first to formalize performative prediction in the supervised-learning setting; their work, along with the majority of subsequent papers [5, 9, 16, 19, 21, 24, 26, 29], have been focused on performative stability and proposed algorithms for learning performative stable parameters. On the other hand, performative optimality requires much stricter conditions (e.g. distributional assumptions to ensure the convexity of $\text{PR}(\theta)$) than performative stability, a few papers address the problem of finding performative optimal parameters. [20] introduce a two-stage method that learns a distribution map to locate the performative optimal parameter. [17] study performative optimality in outcome-only performative settings. Finally, [11]

provide a comprehensive overview of learning algorithms, optimization methods, and applications for performative prediction. Unlike these prior works on performative prediction, which focus on prediction accuracy, our work is dedicated to constructing powerful and statistically valid inference procedures under the performative framework.

Prediction-powered inference. [2] first introduced the prediction-powered inference (PPI) framework, which leverages black-box machine learning models to construct valid confidence intervals (CIs) for statistical quantities. Since then, PPI has been extended and applied in various settings. Closely related to our strategies, [1] propose PPI++, a more computationally efficient procedure that enhances predictability by accommodating a wider range of models on unlabeled data, while guaranteeing performance (e.g. CI width) no worse than that of classical inference methods. Other extensions include Stratified PPI [10], which incorporates simple data stratification strategies into basic PPI estimates; Cross PPI [33], which obtains confidence intervals with significantly lower variability by including model training; Bayesian PPI [12] and FAB-PPI [6], which propose frameworks for PPI based on Bayesian inference. PPI is also connected to topics such as semi-parametric inference and missing-data imputation [8, 27, 28, 30]. Our work is the first one to study PPI under performativity, and we validate the PPI framework in the performative setting both theoretically and empirically.

Inference in performativity. [5] studies identifiability and estimation error under a specific micro-foundation model with performativity. Yet it doesn't address confidence interval construction. A closely related work is [7], which also establishes the asymptotic normality and minimax optimality for performative settings. It focuses on stochastic gradient update with one sample per iteration, whereas our work analyzes the empirical risk minimizer on batch updates. Moreover, [7] does not provide a data-driven approach for covariance estimation and thus lacks an end-to-end inference framework. In contrast, our work explicitly handles density estimation and provides an end-to-end inference method for constructing confidence intervals, which is missing in existing literature, to the best of our knowledge.

2 Background

In this section, we recap more detailed background knowledge about performative prediction and prediction-powered inference.

Repeated risk minimization. Recall that the main objective of interest is the *performative stable* point, which satisfies the following relationship:

$$\theta_{\text{PS}} = \arg \min_{\theta} \mathbb{E}_{z \sim \mathcal{D}(\theta_{\text{PS}})} \ell(z; \theta).$$

Repeated risk minimization (RRM) is a simple algorithm that can efficiently find θ_{PS} . Specifically, one starts with an arbitrary θ_0 and repeat the following procedure:

$$\theta_{t+1} = \arg \min_{\theta} \mathbb{E}_{z \sim \mathcal{D}(\theta_t)} \ell(z; \theta)$$

for $t \in \mathbb{N}$. Under some regularity conditions, the above update is well-defined and provably converges to a unique θ_{PS} at a linear rate.

Theorem 2.1 (Informal, adopted from [25]). *If the loss is smooth, strongly convex, and the mapping $\mathcal{D}(\cdot)$ satisfies certain Lipschitz conditions, then θ_{PS} is uniquely defined and repeated risk minimization converges to θ_{PS} in a linear rate.*

We will further explicitly state those conditions in Section 3. Throughout the paper, we will mainly focus on building an inference framework under the repeated risk minimization algorithm.

Prediction-powered inference. A rich line of work on prediction-powered inference (PPI) [1] considers how to combine limited gold-standard labeled data with abundant unlabeled data to obtain more efficient estimation and construct tighter confidence regions for some unknown parameters. Specifically, a general predictive setting is considered in which each instance has an input $x \in \mathcal{X}$ and an associated observation $y \in \mathcal{Y}$. People have access to a limited set of gold-standard labeled data $\{x_i, y_i\}_{i=1}^n$ that are i.i.d. drawn from a distribution $\mathcal{D}$. Meanwhile, we have abundant unlabeled data $\{x_i^u\}_{i=1}^N$ that are i.i.d. drawn from the same marginal distribution as gold-standard labeled data, i.e. $\mathcal{D}_{\mathcal{X}}$, where $N \gg n$. In addition, an annotating model $f : \mathcal{X} \mapsto \mathcal{Y}$ (possibly off-the-shelf and black-

box machine learning models) is used to label data¹. In [1], the authors show that for a convex loss with a unique solution, compared with standard M-estimator $\hat{\theta}^{\text{SL}} = \arg \min_{\theta} \sum_{i=1}^n \ell(x_i, y_i; \theta)/n$,

$$\hat{\theta}^{\text{PPI}}(\lambda) := \arg \min_{\theta} \lambda \frac{1}{N} \sum_{i=1}^N \ell(x_i^u, f(x_i^u); \theta) + \frac{1}{n} \sum_{i=1}^n \ell(x_i, y_i; \theta) - \lambda \frac{1}{n} \sum_{i=1}^n \ell(x_i, f(x_i); \theta)$$

can be a better estimator of $\theta^* = \arg \min_{\theta} \mathbb{E}_{z \sim \mathcal{D}} \ell(z; \theta)$ via appropriately chosen λ based on data.

Notation. For $K \in \mathbb{N}_+$, we use $[K]$ to denote $\{1, 2, \dots, K\}$. We use $\xrightarrow{P}$ and $\xrightarrow{D}$ to denote convergence in probability and in distribution, respectively. For two set $\mathcal{S}$ and $\mathcal{S}'$, we use $\mathcal{S} + \mathcal{S}'$ to denote the set $\{s + s' : s \in \mathcal{S}, s' \in \mathcal{S}'\}$. $\mathcal{N}(\mu, \Sigma)$ denotes a Gaussian distribution with mean μ and covariance matrix Σ . Lastly, we denote a k -dimensional identity matrix as I_k . We use $\mathcal{B}(c, r)$ to denote a ball with center c and radius r . Lastly, we use $\|\cdot\|$ for ℓ_2 -norm and $\mathbf{1}$ to denote a column vector with all coordinates 1.

3 Inference under Performativity

In this section, we initiate the inference framework for repeated risk minimization in the batch setting under performativity. We mainly consider the repeated risk minimization setting. Unlike standard inference problems, where estimators are built for a *fixed* underlying data distribution, in our *dynamic* setting specified below, building asymptotic results such as CLT imposes extra challenges and this has not been covered by any existing literature so far.

Specifically, at time $t = 0$, we have access to a set of labeled data $\{z_{0,i}\}_{i=1}^{n_0}$ that is i.i.d. drawn from a distribution $\mathcal{D}(\theta_0)$, where $z_{0,i} = (x_{0,i}, y_{0,i})$ and θ_0 is chosen by us. Then, we use the empirical repeated risk minimization to output

$$\hat{\theta}_1 = \arg \min_{\theta} \frac{1}{n_0} \sum_{i=1}^{n_0} \ell(z_{0,i}; \theta)$$

as an estimator of $\theta_1 = \arg \min_{\theta} \mathbb{E}_{z \sim \mathcal{D}(\theta_0)} \ell(z; \theta)$. Then, for $t \geq 1$, at time t , we further have access to a set of labeled data $\{z_{t,i}\}_{i=1}^{n_t}$ that are i.i.d. drawn from the distribution induced by last iteration, i.e., $\mathcal{D}(\hat{\theta}_t)$. Let us further define $G(\hat{\theta}) = \arg \min_{\theta} \mathbb{E}_{z \sim \mathcal{D}(\hat{\theta})} \ell(z; \theta)$. Then, we can obtain the output

$$\hat{\theta}_{t+1} = \arg \min_{\theta} \frac{1}{n_t} \sum_{i=1}^{n_t} \ell(z_{t,i}; \theta)$$

as an estimator of $\theta_{t+1} = G(\hat{\theta}_t)$. This iterative process will incur two trajectories, i.e., (1) **underlying trajectory**: $\theta_0 \rightarrow \theta_1 \rightarrow \dots \theta_t \rightarrow \dots$; (2) **trajectory in practice**: $\theta_0 \rightarrow \hat{\theta}_1 \rightarrow \dots \hat{\theta}_t \rightarrow \dots$.

Our aim is to provide inference on $\hat{\theta}_t$ for any $t \geq 1$. For simplicity, we let $n_t = n$ for all t .

3.1 Central Limit Theorem of $\hat{\theta}_t$

In order to build CLT for $\hat{\theta}_t$, we first establish the consistency of $\hat{\theta}_t$, which is relatively straightforward given that [25] has built the non-asymptotic convergence results. Then, we introduce our main result on building CLT for $\hat{\theta}_t$. Lastly, we provide a novel method to estimate the variance of $\hat{\theta}_t$.

Consistency of $\hat{\theta}_t$. We start with proving the consistency of $\hat{\theta}_t$. Recall that we have a trajectory induced by the samples $\theta_0 \rightarrow \hat{\theta}_1 \rightarrow \dots \hat{\theta}_t \rightarrow \dots$ by the iterative algorithm deployed. Without consistency, CLT is not expected to hold. Our results are based on the following assumptions.

Assumption 3.1. Assume the loss function ℓ satisfies:

- (a). (*Local Lipschitzness*) Loss function $\ell(z; \theta)$ is locally Lipschitz: for each θ , there exist a neighborhood $\Upsilon(\theta)$ of θ such that $\ell(z; \tilde{\theta})$ is $L(z)$ Lipschitz w.r.t $\tilde{\theta}$ for all $\tilde{\theta} \in \Upsilon(\theta)$ and $\mathbb{E}_{z \sim \mathcal{D}(\theta)} L(z) < \infty$.

¹The annotating function f could either be a stochastic or deterministic function. It could even take other inputs besides x , but for simplicity, we only consider the annotation with the form $f(x)$.

(b). (*Joint Smoothness*) Loss function $\ell(z; \theta)$ is β -jointly smooth in both z and θ :

$$\|\nabla_{\theta}\ell(z; \theta) - \nabla_{\theta}\ell(z; \theta')\|_2 \leq \beta \|\theta - \theta'\|_2, \quad \|\nabla_{\theta}\ell(z; \theta) - \nabla_{\theta}\ell(z'; \theta)\|_2 \leq \beta \|z - z'\|_2,$$

for any $z, z' \in \mathcal{Z}$ and $\theta, \theta' \in \Theta$.

(c). (*Strong Convexity*) Loss function $\ell(z; \theta)$ is γ -strongly convex w.r.t θ :

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

for any $z \in \mathcal{Z}$ and θ, θ' .

(d). (ε -Sensitivity) The distribution map $D(\theta)$ is ε -sensitive, i.e.:

$$W_1(D(\theta), D(\theta')) \leq \varepsilon \|\theta - \theta'\|,$$

for any θ, θ' , where W_1 is the Wasserstein-1 distance.

Remark 3.2. The assumptions (b), (c), (d) follow the standard ones in [25], which are proved to be the minimal requirements for trajectory convergence. We additionally require (a) to build consistency for $\hat{\theta}_t$ beyond convergence of $\hat{\theta}_t$ to θ_{PS} .

Proposition 3.3. Under Assumption 3.1, if $\varepsilon < \frac{\gamma}{\beta}$, then for any given $T \geq 0$, we have that for all $t \in [T]$,

$$\hat{\theta}_t \xrightarrow{P} \theta_t.$$

Building CLT for $\hat{\theta}_t$. In order to build the central limit theorem for $\hat{\theta}_t$, we need to introduce a few extra assumptions. Due to limited space, going forward, we defer the required assumptions in later theorems to Appendix.

Let us denote $\Sigma_{\tilde{\theta}}(\theta) = H_{\tilde{\theta}}(\theta)^{-1} V_{\tilde{\theta}}(\theta) H_{\tilde{\theta}}(\theta)^{-1}$, where $H_{\tilde{\theta}}(\theta) = \nabla_{\tilde{\theta}}^2 \mathbb{E}_{z \sim \mathcal{D}(\tilde{\theta})} \ell(z; \theta)$ and $V_{\tilde{\theta}}(\theta) = \text{Cov}_{z \sim \mathcal{D}(\tilde{\theta})}(\nabla_{\theta} \ell(z; \theta))$. And recall that $G(\tilde{\theta}) = \arg \min_{\theta} \mathbb{E}_{z \sim \mathcal{D}(\tilde{\theta})} \ell(z; \theta)$.

Theorem 3.4 (Central Limit Theorem of $\hat{\theta}_t$). Under Assumption 3.1 and A.1, if $\varepsilon < \frac{\gamma}{\beta}$, then for any given $T \geq 0$, we have that for all $t \in [T]$,

$$\sqrt{n}(\hat{\theta}_t - \theta_t) \xrightarrow{D} \mathcal{N}(0, V_t)$$

with

$$V_t = \sum_{i=1}^t \left[\prod_{k=i}^{t-1} \nabla G(\theta_k) \right] \Sigma_{\theta_{i-1}}(\theta_i) \left[\prod_{k=i}^{t-1} \nabla G(\theta_k) \right]^{\top}.$$

In particular, $\nabla G(\theta_k) = -H_{\theta_k}(\theta_{k+1})^{-1} (\nabla_{\tilde{\theta}} \mathbb{E}_{z \sim \mathcal{D}(\theta_k)} \nabla_{\theta} \ell(z; \theta_{k+1}))$, where $\nabla_{\tilde{\theta}}$ is taking gradient for the parameter in $\mathcal{D}(\tilde{\theta})$, ∇_{θ} is taking gradient for the parameter in $\ell(z; \theta)$ and $\prod_{k=t}^{t-1} \nabla G(\theta_k) = I_d$.

Estimation of $\nabla G(\theta_t)$, V_t Given the established CLT for $\hat{\theta}_t$, in order to construct confidence regions for $\hat{\theta}_t$ in practice, the only thing left is to provide an estimation of $\nabla G(\theta_t)$ and V_t with samples. In previous results, with a more detailed calculation, we obtain

$$\nabla G(\theta_k) = -H_{\theta_k}(\theta_{k+1})^{-1} \mathbb{E}_{z \sim \mathcal{D}(\theta_k)} [\nabla_{\theta} \ell(z, \theta_{k+1}) \nabla_{\theta} \log p(z, \theta_k)^{\top}].$$

where $p(\cdot, \theta)$ is the density function of distribution $D(\theta)$, and the score function $\nabla_{\theta} \log p(z, \theta)$ is thus a d -dimensional vector (recall that θ is of dimension d). In order to estimate it for any θ , we propose a novel score matching method. Specifically, we use a model $M(z, \theta; \psi)$ parameterized by ψ to approximate $p(z, \theta)$. Inspired by the objective in [14], for any given θ (e.g., $\hat{\theta}_t$), we aim to minimize the following objective parameterized by ψ **for all** θ :

$$\begin{aligned} J(\theta; \psi) &= \int p(z, \theta) \|\nabla_{\theta} \log p(z, \theta) - s(z, \theta; \psi)\|^2 dz \\ &= \int p(z, \theta) \left(\|\nabla_{\theta} \log p(z, \theta)\|^2 + \|s(z, \theta; \psi)\|^2 - 2 \nabla_{\theta} \log p(z, \theta)^{\top} s(z, \theta; \psi) \right) dz \end{aligned}$$

where $s(z, \theta; \psi) = \nabla_{\theta} \log M(z, \theta; \psi)$. If we can learn a $\hat{\psi}$ so that $s(z, \theta; \hat{\psi}) = \nabla_{\theta} \log p(z, \theta)$ for all θ , then we can reach the minimum $J(\theta; \hat{\psi}) = 0$.

Notice the first term is unrelated to ψ ; the second term involves the model M that is chosen by us, so we have the analytical expression of it. Thus, our key task will be estimating the third term, which involves $\mathcal{K}(\theta; \psi) := \int p(z, \theta) \nabla_{\theta} \log p(z, \theta)^{\top} s(z, \theta; \psi) dz$.

We remark that in our setting, instead of taking the gradient at z , we have new challenges in taking the gradient at θ . So, we derive the following key lemma.

Lemma 3.5. *Under Assumption A.2, we have*

$$\mathcal{K}(\theta; \psi) = \sum_{i=1}^d \left[\frac{\partial}{\partial \theta^{(i)}} \int p(z, \theta) \frac{\partial \log M(z, \theta; \psi)}{\partial \theta^{(i)}} dz - \int p(z, \theta) \frac{\partial^2 \log M(z, \theta; \psi)}{\partial \theta^{(i)2}} dz \right]$$

where $\theta^{(i)}$ is the i -th coordinate of θ .

Based on the lemma, we propose a novel gradient-free score matching method with *policy perturbation* to estimate $\mathcal{K}(\hat{\theta}_t; \psi)$ for any $t \in [T]$. Policy perturbation is a commonly used technique in estimating the policy effect under general equilibrium shift [22] or interference [32]. Instead of just getting samples for $\hat{\theta}_t$ for each t , we additionally sample for all perturbed policies in $\{\hat{\theta}_t + \eta e_1, \hat{\theta}_t + \eta e_2, \dots, \hat{\theta}_t + \eta e_d\}$, where $\eta > 0$ is a small scalar at our choice and $\{e_j\}_j$ are standard basis for $\mathbb{R}^d$. Typically, for a policy θ , its dimension d is low. One concrete example in practice is to use slightly different price strategies in different local markets.

Specifically, the term $\int p(z, \hat{\theta}_t) \frac{\partial^2 \log M(z, \hat{\theta}_t; \psi)}{\partial \theta^{(i)2}} dz$ could be easily estimated by using empirical mean, e.g., $\frac{1}{n} \sum_{j=1}^n \frac{\partial^2 \log M(z_{t,j}, \hat{\theta}_t; \psi)}{\partial \theta^{(i)2}}$. And for the derivative $\frac{\partial}{\partial \theta^{(i)}} \int p(z_{t,j}, \hat{\theta}_t) \frac{\partial \log M(z_{t,j}, \hat{\theta}_t; \psi)}{\partial \theta^{(i)}} dz$, if we draw additional k samples $\{z_{t,u}^{(i)}\}_{u=1}^k$ for each perturbed policy $\hat{\theta}_t + \eta e_i$, we can use the following estimator:

$$\frac{1}{\eta} \left(\frac{1}{k} \sum_{u=1}^k \frac{\partial \log M(z_{t,u}^{(i)}, \hat{\theta}_t + \eta e_i; \psi)}{\partial \theta^{(i)}} - \frac{1}{n} \sum_{u=1}^n \frac{\partial \log M(z_{t,u}, \hat{\theta}_t; \psi)}{\partial \theta^{(i)}} \right).$$

Combining the above, we have a straightforward way to estimate $G(\theta_t)$ and V_t for any $t \in [T]$ by plugging in the empirical estimate. Let us denote the estimator of V_t by $\hat{V}_t$. Then, we would have

$$\hat{V}_t^{-1/2} \sqrt{n}(\hat{\theta}_t - \theta_t) \xrightarrow{D} \mathcal{N}(0, I_d) \quad (2)$$

if the model $M(z, \theta; \psi)$ is expressive enough.

Theoretical validity of estimation. Now we prove that the policy perturbation method provides a valid estimator of $\nabla G(\theta_k)$. Recall that the gradient of G is given by

$$\nabla G(\theta_k) = -H_{\theta_k}(\theta_{k+1})^{-1} \mathbb{E}_{\theta_k} [\nabla_{\theta} \ell(z; \theta_{k+1}) \nabla_{\theta} \log p(z, \theta_k)^{\top}],$$

and the estimator is defined by

$$\hat{g}_k := -H_{\hat{\theta}_k}(\hat{\theta}_{k+1})^{-1} \hat{\mathbb{E}}_{\hat{\theta}_k} [\nabla_{\theta} \ell(z; \hat{\theta}_{k+1}) s(z, \hat{\theta}_k; \hat{\psi}(\hat{\theta}_k))^{\top}].$$

Theorem 3.6. *Under Assumption 3.1 A.1, A.9 and A.11, we have*

$$\|\hat{g}_k - \nabla G(\theta_k)\|^2 = O_p\left(\frac{1}{\sqrt{n}} + \frac{1}{\eta \sqrt{\min(n, k)}} + \eta + a_n\right).$$

where $a_n = o(1)$ is a vanishing optimization error term defined in Assumption A.9(d).

The proof is by decomposing the error between the empirical and true gradient into several components, where each varies in whether it uses empirical or population quantities. We bound the deviations between the true parameters θ_k, θ_{k+1} and their empirical counterparts $\hat{\theta}_k, \hat{\theta}_{k+1}$ using Theorem 3.4. We bound the true score function and its estimation via a uniform bound of $\|\nabla \log p(z, \theta) - s(z, \theta, \psi(\theta))\|^2$ over the perturbed policies using Theorem A.13. Taken together, the above result shows the consistency of $\hat{g}_k$ obtained by policy perturbation.

3.2 Bias-Aware Inference for Performative Stable Point

Finally, we further provide a way to construct the confidence region for θ_{PS} . This is directly followed from our previous results on building CLT for $\hat{\theta}_t$. By the convergence results derived for the underlying trajectory by [25], under Assumption 3.1, we have

$$\|\theta_t - \theta_{\text{PS}}\| \leq \left(\frac{\varepsilon\beta}{\gamma}\right)^t \|\theta_0 - \theta_{\text{PS}}\|.$$

Thus, we can immediately obtain the following corollary by using bias-aware inference – a commonly seen technique in econometrics [3, 4, 15, 23].

Corollary 3.7 (Confidence region construction for θ_{PS}). *Under Assumption 3.1, A.1, A.2, and A.3, if $\varepsilon < \frac{\gamma}{\beta}$, for any $\delta \in (0, 1)$, we can obtain a confidence region $\hat{\mathcal{R}}_t(n, \delta)$ for θ_t by using Eq. 2, such that*

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\theta_t \in \hat{\mathcal{R}}_t(n, \delta)\right) = 1 - \delta.$$

Moreover, if $\theta_0, \theta_{\text{PS}} \in \{\theta : \|\theta\| \leq B\}$,

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\theta_{\text{PS}} \in \hat{\mathcal{R}}_t(n, \delta) + \mathcal{B}\left(0, 2B\left(\frac{\varepsilon\beta}{\gamma}\right)^t\right)\right) \geq 1 - \delta.$$

Corollary 3.7 provides a way to construct the confidence region for the performative stable point based on the confidence region for θ_t . Notice that the derived new confidence region is quite close to $\hat{\mathcal{R}}_t(n, \delta)$ and the difference vanishes exponentially fast as t grows. Thus, we expect the derived region to be quite tight for moderately large t , meaning:

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\theta_{\text{PS}} \in \hat{\mathcal{R}}_t(n, \delta) + \mathcal{B}\left(0, 2B\left(\frac{\varepsilon\beta}{\gamma}\right)^t\right)\right) \approx 1 - \delta.$$

For the condition $\theta_0, \theta_{\text{PS}} \in \{\theta : \|\theta\| \leq B\}$, it will be natural to satisfy and we can get an explicit and feasible upper bound B under mild conditions. It is because that θ_0 is at our choice and we can further derive an explicit and feasible upper bound for $\|\theta_{\text{PS}}\|$ by using the strong convexity. Specifically, by γ -strong convexity of the loss function with respect to θ , we have

$$\left(\mathbb{E}_{z \sim \mathcal{D}(\theta_{\text{PS}})} \nabla_{\theta} \ell(z; \theta_0) - \mathbb{E}_{z \sim \mathcal{D}(\theta_{\text{PS}})} \nabla_{\theta} \ell(z; \theta_{\text{PS}})\right)^{\top} (\theta_0 - \theta_{\text{PS}}) \geq \gamma \|\theta_0 - \theta_{\text{PS}}\|^2.$$

Since $\mathbb{E}_{z \sim \mathcal{D}(\theta_{\text{PS}})} \nabla_{\theta} \ell(z; \theta_{\text{PS}}) = 0$, this leads to $\|\mathbb{E}_{z \sim \mathcal{D}(\theta_{\text{PS}})} \nabla_{\theta} \ell(z; \theta_0)\| \geq \gamma \|\theta_0 - \theta_{\text{PS}}\|$. Thus, if we further have $\sup_{z \in \mathcal{Z}} \|\nabla_{\theta} \ell(z; \theta_0)\| \leq \tilde{B}$ for $\tilde{B} > 0$, which could be achieved and calculated by assuming $\mathcal{Z}$ is compact and use the continuity of $\nabla_{\theta} \ell(\cdot, \theta_0)$. Then, it will immediately give us an upper bound for $\|\theta_{\text{PS}}\|$ that $\|\theta_{\text{PS}}\| \leq \tilde{B}/\gamma + \|\theta_0\|$.

4 Prediction-Powered Inference under Performativity

In this section, we further investigate prediction-powered inference (PPI) under performativity to enhance estimation and obtain improved confidence regions for the model parameter (i.e., policy) under performativity. This can also address the data scarcity issue in human responses when doing a survey to get feedback on policy implementation.

Specifically, at time $t = 0$, besides the limited set of gold-standard labeled data $\{x_{0,i}, y_{0,i}\}_{i=1}^{n_0}$ that are i.i.d. drawn from a distribution $\mathcal{D}(\theta_0)$, we have abundant unlabeled data $\{x_{0,i}^u\}_{i=1}^{N_0}$ that are i.i.d. drawn from the same marginal distribution as gold-standard labeled data, i.e. $\mathcal{D}_{\mathcal{X}}(\theta_0)$, where $N_0 \gg n_0$. In addition, an annotating model $f : \mathcal{X} \mapsto \mathcal{Y}$ is used to label data², which leads to $\{x_{0,i}, f(x_{0,i})\}_{i=1}^{n_0}$ and $\{x_{0,i}^u, f(x_{0,i}^u)\}_{i=1}^{N_0}$. Then, we use the following mechanism to output

$$\hat{\theta}_1^{\text{PPI}}(\lambda_1) = \arg \min_{\theta} \frac{\lambda_1}{N} \sum_{i=1}^N \ell(x_{0,i}^u, f(x_{0,i}^u); \theta) + \frac{1}{n} \sum_{i=1}^n (\ell(x_{0,i}, y_{0,i}; \theta) - \lambda_1 \ell(x_{0,i}, f(x_{0,i}); \theta))$$

²Our theory can easily be extended to allow using different annotating function for each iteration, but for simplicity in presentation, we use f for all iterations.

for a scalar λ_1 as an estimator of θ_1 . After that, for $t \geq 1$, at time t , besides having access to the set of gold-standard labeled data $\{x_{t,i}, y_{t,i}\}_{i=1}^{n_t}$ that are i.i.d. drawn from a distribution $\mathcal{D}(\hat{\theta}_t)$. Meanwhile, we have abundant unlabeled data $\{x_{t,i}^u\}_{i=1}^{N_t}$ that are i.i.d. drawn from the same marginal distribution as gold-standard labeled data, i.e. $\mathcal{D}_{\mathcal{X}}(\hat{\theta}_t)$, where $N_t \gg n_t$. Similar as before, we can estimate θ_{t+1} via

$$\hat{\theta}_{t+1}^{\text{PPI}}(\lambda_{t+1}) = \arg \min_{\theta} \frac{\lambda_{t+1}}{N} \sum_{i=1}^N \ell(x_{t,i}^u, f(x_{t,i}^u); \theta) + \frac{1}{n} \sum_{i=1}^n (\ell(x_{t,i}, y_{t,i}; \theta) - \lambda_{t+1} \ell(x_{t,i}, f(x_{t,i}); \theta))$$

for a scalar λ_{t+1} . This incurs a trajectory in practice: $\theta_0 \rightarrow \hat{\theta}_1^{\text{PPI}}(\lambda_1) \rightarrow \dots \hat{\theta}_t^{\text{PPI}}(\lambda_t) \rightarrow \dots$. Notice that if we choose $\lambda_t = 0$ for all $t \in [T]$, this will degenerate to the case in Section 3. Later on, we will demonstrate how to choose $\{\lambda_t\}_{t=1}^T$ via data to enhance inference. Our mechanism could be adaptive to the data quality with carefully chosen $\{\lambda_t\}_{t=1}^T$ and could be viewed as an extension of the classical PPI++ mechanism [25] to the setting under performativity.

Building CLT for $\hat{\theta}_t^{\text{PPI}}(\lambda_t)$. We start with building the central limit theorem for $\hat{\theta}_t^{\text{PPI}}(\lambda_t)$ with fixed constant scalars $\{\lambda_t\}_{t=1}^T$ for any $t \in [T]$. The proof is similar to that of Theorem 3.4. Specifically, we denote

$$\Sigma_{\lambda, \bar{\theta}}(\theta; r) = H_{\bar{\theta}}(\theta)^{-1} \left(r V_{\lambda, \bar{\theta}}^f(\theta) + V_{\lambda, \bar{\theta}}(\theta) \right) H_{\bar{\theta}}(\theta)^{-1}$$

with $V_{\lambda, \bar{\theta}}^f(\theta) = \lambda^2 \text{Cov}_{x \sim \mathcal{D}_{\mathcal{X}}(\bar{\theta})}(\nabla_{\theta} \ell(x, f(x); \theta))$ and $V_{\lambda, \bar{\theta}}(\theta) = \text{Cov}_{(x,y) \sim \mathcal{D}(\bar{\theta})}(\nabla_{\theta} \ell(x, y; \theta) - \lambda \nabla_{\theta} \ell(x, f(x); \theta))$. Then, we have the following theorem.

Theorem 4.1 (Central Limit Theorem of $\hat{\theta}_t^{\text{PPI}}(\lambda_t)$). *Under Assumption 3.1, A.4, and A.5, if $\varepsilon < \frac{\gamma}{\beta}$ and $\frac{n}{N} \rightarrow r$ for some $r \geq 0$, then for any given $T \geq 0$, we have that for all $t \in [T]$,*

$$\sqrt{n}(\hat{\theta}_t^{\text{PPI}}(\lambda_t) - \theta_t) \xrightarrow{D} \mathcal{N}\left(0, V_t^{\text{PPI}}(\{\lambda_j, \theta_j\}_{j=1}^t; r)\right)$$

with

$$V_t^{\text{PPI}}(\{\lambda_j, \theta_j\}_{j=1}^t; r) = \sum_{i=1}^t \left[\prod_{k=i}^{t-1} \nabla G(\theta_k) \right] \Sigma_{\lambda_i, \theta_{i-1}}(\theta_i; r) \left[\prod_{k=i}^{t-1} \nabla G(\theta_k) \right]^{\top}.$$

Selection of parameters $\{\lambda_t\}_{t=1}^T$. Now, the only thing left is to select $\{\lambda_t\}_{t=1}^T$, so as to enhance estimation and inference. As choosing $\lambda_t = 0$ for all $t \in [T]$ will degenerate to Theorem 3.4, we expect that we appropriately choose $\{\lambda_t\}_{t=1}^T$ to make $F(V_t) \geq F(V_t^{\text{PPI}}(\{\lambda_j, \theta_j\}_{j=1}^t; r))$, where F is a user-specified scalarization operator depending on different aims. For instance, if we are interested in optimizing the sum of asymptotic variance of coordinates of θ_t , then, $F(V_t) = \text{Tr}(V_t)$. Or if we are interested in the inference of the sum of all coordinates θ_t , i.e., $\mathbf{1}^{\top} \theta_t$, then $F(V_t) = \mathbf{1}^{\top} V_t \mathbf{1}$.

We consider a greedy sequential selection mechanism as follows. Imagine that we have selected $\{\lambda_i^*\}_{i=1}^{t-1}$ via the data, and our aim is to select a λ_t^* so as to make $F(V_t^{\text{PPI}}(\{\lambda_j^*, \theta_j\}_{j=1}^{t-1}, \lambda, \theta_t; r)) \geq F(V_t^{\text{PPI}}(\{\lambda_j^*, \theta_j\}_{j=1}^t; r))$ for any λ . Thus, we choose

$$\lambda_t^* = \arg \min_{\lambda} F\left(V_t^{\text{PPI}}(\{\lambda_j^*, \theta_j\}_{j=1}^{t-1}, \lambda, \theta_t; r)\right). \quad (3)$$

However, in practice, there are still several issues that need to be addressed. First, we need to choose $\{\lambda_t\}_{t=1}^T$ via observations, and this could be handled by using our results in Section 3 to estimate $\nabla_{\theta} G(\theta)$, and we obtain a sample version $\hat{\lambda}_t$ by plugging in the estimation. Second, when obtaining λ_t^* in Eq. 3, we actually need θ_t in $\Sigma_{\lambda_i, \theta_{i-1}}(\theta_i; r)$. But when using samples to estimate, we need to get $\hat{\lambda}_t$ first before we obtain $\hat{\theta}_t^{\text{PPI}}$. Thus, we propose a similar optimizing strategy as inspired by [1]: at time $t \geq 1$, given the obtained $\{\hat{\lambda}_i\}_{i=1}^{t-1}$ and $\{\hat{\theta}_i^{\text{PPI}}\}_{i=1}^{t-1}$, we choose an arbitrary $\tilde{\lambda}$, to obtain $\hat{\theta}_t^{\text{PPI}}(\tilde{\lambda})$ as a temporary surrogate³. Then, we further obtain

$$\hat{\lambda}_t = \arg \min_{\lambda} F\left(\hat{V}_t^{\text{PPI}}(\{\hat{\lambda}_j, \hat{\theta}_j^{\text{PPI}}(\hat{\lambda}_j)\}_{j=1}^{t-1}, \lambda, \hat{\theta}_t^{\text{PPI}}(\tilde{\lambda}); \frac{n}{N}\right),$$

³Notice that $\hat{\theta}_t^{\text{PPI}}(\tilde{\lambda})$ is still a consistent estimator of θ_t

where $\widehat{V}_t^{\text{PPI}}$ is obtained via replacing $\nabla_{\theta} G(\theta_k)$ with their estimation in V_t^{PPI} . In particular, if F satisfies $F(U + V) = F(U) + F(V)$ as our examples of $F(V_t) = \text{Tr}(V_t)$ and $F(V_t) = \mathbf{1}^\top V_t \mathbf{1}$, then we only need to optimize $F\left(\Sigma_{\lambda, \widehat{\theta}_{t-1}^{\text{PPI}}(\widehat{\lambda}_{t-1})}(\widehat{\theta}_t(\widehat{\lambda}); \frac{n}{N})\right)$

To sum up, by the above process, we have the following corollary.

Corollary 4.2. *Under Assumption 3.1, A.2, A.3, A.4, and A.5, if $\varepsilon < \frac{\gamma}{\beta}$ and $\frac{n}{N} \rightarrow r$ for some $r \geq 0$, then for any given $T \geq 0$, we have that for all $t \in [T]$,*

$$\left(\widehat{V}_t^{\text{PPI}}\left(\{\widehat{\lambda}_j, \widehat{\theta}_j^{\text{PPI}}(\widehat{\lambda}_j)\}_{j=1}^t; \frac{n}{N}\right)\right)^{-\frac{1}{2}} \sqrt{n}(\widehat{\theta}_t^{\text{PPI}}(\widehat{\lambda}_t) - \theta_t) \xrightarrow{D} \mathcal{N}(0, I_d).$$

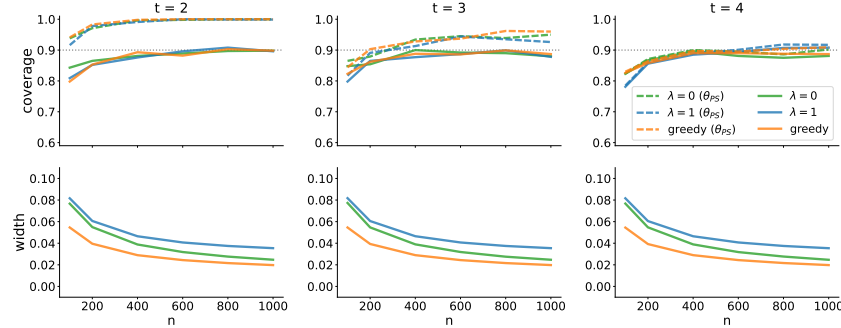


Figure 1: Confidence-region coverage (top row) and width (bottom row) with different choices of λ . The left, middle, and right columns correspond to inference steps $t = 2$, $t = 3$, and $t = 4$, respectively. The solid and dashed curves correspond to the confidence-region coverage for θ_t and θ_{PS} , respectively.

5 Experiment

In this section, we further provide numerical experimental results to support our previous theory. A case study on a semi-synthetic dataset is provided in Section B.3.

Experimental setting. We follow [20] to construct simulation studies on a performative linear regression problem. Given a parameter $\theta \in \mathbb{R}^d$, data are sampled from $D(\theta)$ as

$$y = \alpha^\top x + \mu^\top \theta + \nu, \quad x \sim \mathcal{N}(\mu_x, \Sigma_x), \quad \nu \sim \mathcal{N}(0, \sigma_y^2).$$

The distribution map $D(\theta)$ is ε -sensitive with $\varepsilon = \|\mu\|_2$. For unlabeled x_i^u , the annotating model is defined as $f(x_i^u) = \alpha^\top x_i^u + \mu^\top \theta + \nu_i, \nu_i \sim \mathcal{N}(-0.2, \sigma_y^2)$. We use the ridge squared loss to measure the performance and update θ : $\ell((x, y); \theta) = \frac{1}{2}(y - \theta^\top x)^2 + \frac{\gamma}{2}\|\theta\|^2$. For easier calculation for smoothness parameter, we truncate the the distribution of (x, y) in our experiment to deal with truncated normal distributions, but this is not necessary in many other choices of updating rules. In the following experiments, we set $d = 2$, $\varepsilon \approx 0.02$, $\gamma = 2$, and $\sigma_y^2 = 0.2$. We set $N = 2000$ and vary the labeled sample size n .

Simulation results for PPI under performativity. To quantify the results of PPI under performativity, we evaluate the confidence-region coverage and width for three strategies: $\lambda = 0$ (only labeled data), $\lambda = 1$ (full unlabeled data weight), and our optimization method $\lambda = \widehat{\lambda}_t$ as defined in Eq.3. We vary the labeled sample size n and perform $t \in \{2, 3, 4\}$ repeated risk minimization steps, averaging results over 1000 independent trials. In Figure 1, we can find that all three methods approach 0.9 coverage as n grows, while our optimized $\widehat{\lambda}_t$ (orange) achieves the narrowest interval width, supporting its effectiveness to enhance the performative inference. The dashed curves denote the bias-adjusted confidence regions for the performative stable point θ_{PS} . It can be observed that θ_{PS} coverages upper-bound that of θ_t (solid curves) across steps t , and the gap between them vanishes as t grows. This observation verifies the validity of Corollary 3.7.

Verifying central limit theorem. To validate the central limit theorem, we sample different $\widehat{\theta}_t$ (here $t = 4$ and $n = 1000$) and visualize the distribution of $\widehat{V}_t^{-1/2} \sqrt{n}(\widehat{\theta}_t - \theta_t)$. We plot the density map

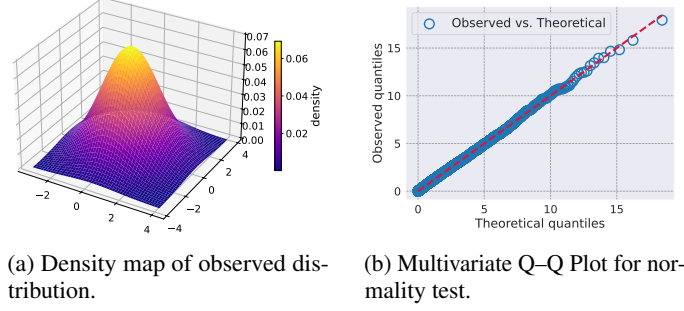


Figure 2: Visualizations to verify the Central Limit Theorem. (a) plots the density map of sampled $\hat{V}_t^{-1/2} \sqrt{n}(\hat{\theta}_t - \theta_t)$, while (b) compares the observed distribution with theoretical one $\mathcal{N}(0, I_d)$.

in Figure 2a and find it is close to a normal distribution. In Figure 2b, we further do a normality test with the multivariate Q-Q plot of observed squared Mahalanobis distance over theoretical Chisq quantiles. The tight alignment of points along the identity line (red dashed) verifies that the observed distribution is well-approximated by its asymptotic CLT.

Estimation results of score matching. We consider two implementations of the gradient-free score matching estimator $M(z, \theta; \psi)$:

- (a). Gaussian parametric: we assume $p(z, \theta) = \mathcal{N}(\mu_p, \Sigma_p)$ and parameterize $\psi = \{\mu_p, \Sigma_p\}$;
- (b). DNN-based: a small deep neural network with two hidden layers of width 128.

We collect $\hat{\theta}_{1:t}$ trajectory and corresponding data $\{z_{1:t,i}\}_{i=1}^n$ to train both models via the SGD optimizer with a learning rate of 0.1 to minimize the empirical score-matching objective $J(\psi)$.

In Figure 3, we evaluate the estimation quality of two models by their final training loss $J(\psi)$ and the estimated variance error $\|\hat{V}_t - V_t\|$ over varying n and t . We sample 1000 independent trajectories of $\hat{\theta}_{1:t}$ and report $J(\psi)$ as averaged $J(\hat{\theta}; \psi)$ over all $\hat{\theta}$ in the collection of trajectories. In all settings, both estimators achieve $J(\psi) < 0.05$, indicating the perfect approximation of our learned model $M(z, \theta; \psi)$ to the true $p(z, \theta)$. Correspondingly, the variance-estimation error remains negligible and decreases as n grows, verifying the feasibility of using our score matching models to fit $\nabla_{\theta} \log p(z, \theta)$ for estimating $\nabla G(\theta_k)$ and V_t .

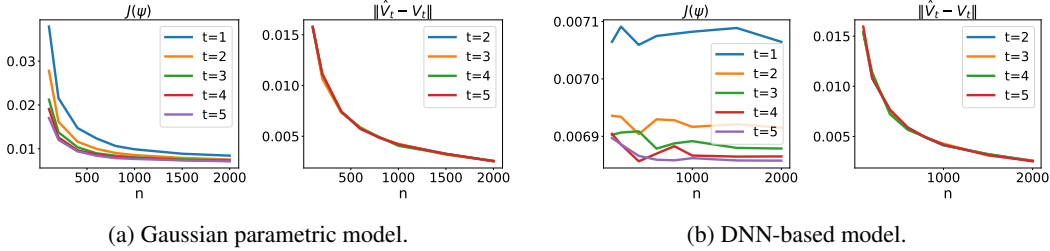


Figure 3: Evaluating the estimation quality of two designed score matching models.

6 Conclusion

In this paper, we introduce an important topic: statistical inference under performativity. We derive results on asymptotic distributions for a widely used iterative process for updating parameters in performative prediction. We further leverage and extend prediction-powered inference to the dynamic setting under performativity. Currently, our framework uses bias-awared inference for θ_{ps} . Obtaining direct inference methods will be of future interest. Our work can serve as an important tool and guideline for policy-making in a wide range of areas such as social science and economics.

7 Acknowledgements

We would like to acknowledge the helpful discussion and suggestions from Tijana Zrinc, Juan Perdomo, Anastasios Angelopoulos, and Steven Wu.

References

- [1] Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023.
- [2] Anastasios Nikolas Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382:669 – 674, 2023. URL <https://api.semanticscholar.org/CorpusID:256105365>.
- [3] Timothy B Armstrong and Michal Kolesár. Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683, 2018.
- [4] Timothy B Armstrong, Michal Kolesár, and Soonwoo Kwon. Bias-aware inference in regularized regression models. *arXiv preprint arXiv:2012.14823*, 2020.
- [5] Daniele Bracale, Subha Maity, Felipe Maia Polo, Seamus Somerstep, Moulinath Banerjee, and Yuekai Sun. Microfoundation inference for strategic prediction. *arXiv preprint arXiv:2411.08998*, 2024.
- [6] Stefano Cortinovis and Francois Caron. Fab-ppi: Frequentist, assisted by bayes, prediction-powered inference. *ArXiv*, abs/2502.02363, 2025. URL <https://api.semanticscholar.org/CorpusID:276107406>.
- [7] Joshua Cutler, Mateo Diaz, and Dmitriy Drusvyatskiy. Stochastic approximation with decision-dependent distributions: asymptotic normality and optimality. *Journal of Machine Learning Research*, 25(90):1–49, 2024.
- [8] Hakan Demirtas. Flexible imputation of missing data. *Journal of Statistical Software*, 85:1–5, 2018. URL <https://api.semanticscholar.org/CorpusID:65237735>.
- [9] Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. *Math. Oper. Res.*, 48:954–998, 2020. URL <https://api.semanticscholar.org/CorpusID:227127621>.
- [10] Adam Fisch, Joshua Maynez, R. Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William W. Cohen. Stratified prediction-powered inference for hybrid language model evaluation. *ArXiv*, abs/2406.04291, 2024. URL <https://api.semanticscholar.org/CorpusID:270285671>.
- [11] Moritz Hardt and Celestine Mendler-Dünner. Performative prediction: Past and future. *ArXiv*, abs/2310.16608, 2023. URL <https://api.semanticscholar.org/CorpusID:264451701>.
- [12] R. Alex Hofer, Joshua Maynez, Bhuwan Dhingra, Adam Fisch, Amir Globerson, and William W. Cohen. Bayesian prediction-powered inference. *ArXiv*, abs/2405.06034, 2024. URL <https://api.semanticscholar.org/CorpusID:269740945>.
- [13] Jason L Huang, Mengqiao Liu, and Nathan A Bowling. Insufficient effort responding: examining an insidious confound in survey data. *Journal of Applied Psychology*, 100(3):828, 2015.
- [14] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.
- [15] Guido Imbens and Stefan Wager. Optimized regression discontinuity designs. *Review of Economics and Statistics*, 101(2):264–278, 2019.
- [16] Pedram J. Khorsandi, Rushil Gupta, Mehrnaz Mofakhami, Simon Lacoste-Julien, and Gauthier Gidel. Tight lower bounds and improved convergence in performative prediction. *ArXiv*, abs/2412.03671, 2024. URL <https://api.semanticscholar.org/CorpusID:274514654>.
- [17] Michael P. Kim and Juan C. Perdomo. Making decisions under outcome performativity. *ArXiv*, abs/2210.01745, 2022. URL <https://api.semanticscholar.org/CorpusID:252693241>.

- [18] Achim Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013.
- [19] Celestine Mendler-Dünnér, Juan C. Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. *ArXiv*, abs/2006.06887, 2020. URL <https://api.semanticscholar.org/CorpusID:219636100>.
- [20] John Miller, Juan C. Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, 2021. URL <https://api.semanticscholar.org/CorpusID:231942295>.
- [21] Mehrnaz Mofakhami, Ioannis Mitliagkas, and Gauthier Gidel. Performative prediction with neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2023. URL <https://api.semanticscholar.org/CorpusID:253180829>.
- [22] Evan Munro, Stefan Wager, and Kuang Xu. Treatment effects in market equilibrium. *arXiv preprint arXiv:2109.11647*, 5, 2021.
- [23] Claudia Noack and Christoph Rothe. Bias-aware inference in fuzzy regression discontinuity designs. *arXiv preprint arXiv:1906.04631*, 2019.
- [24] Caspar Oesterheld, Johannes Treutlein, Emery Cooper, and Rubi Hudson. Incentivizing honest performative predictions with proper scoring rules. *ArXiv*, abs/2305.17601, 2023. URL <https://api.semanticscholar.org/CorpusID:258960363>.
- [25] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünnér, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.
- [26] Juan C. Perdomo. Revisiting the predictability of performative, social events. *ArXiv*, abs/2503.11713, 2025. URL <https://api.semanticscholar.org/CorpusID:277066781>.
- [27] James M. Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90:122–129, 1995. URL <https://api.semanticscholar.org/CorpusID:121261196>.
- [28] Shanshan Song, Yuanyuan Lin, and Yong Zhou. A general m-estimation theory in semi-supervised framework. *Journal of the American Statistical Association*, 119:1065 – 1075, 2023. URL <https://api.semanticscholar.org/CorpusID:257266562>.
- [29] Rohan Taori and Tatsunori Hashimoto. Data feedback loops: Model-driven amplification of dataset biases. In *International Conference on Machine Learning*, 2022. URL <https://api.semanticscholar.org/CorpusID:252118610>.
- [30] Anastasios A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, New York, 2006. ISBN 978-0-387-35458-3. URL <https://doi.org/10.1007/0-387-37345-4>.
- [31] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.
- [32] Davide Viviano and Jess Rudder. Policy design in experiments with unknown interference. *arXiv preprint arXiv:2011.08174*, 4, 2024.
- [33] Tijana Zrnic and Emmanuel J. Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences of the United States of America*, 121, 2023. URL <https://api.semanticscholar.org/CorpusID:263134612>.

A Theoretical Details

We will include omitted technical details in the main context. We first summarize all the required additional assumptions in Section A.1. Then, we provide omitted proofs for Section 3 in Section A.2 and omitted proofs for Section 4 in Section A.4.

A.1 Assumptions

In this subsection, we summarize all the additional assumptions we will use to build various theoretical results in this paper. Before we state the assumptions, we need some further notations.

Let us denote $\Psi(\theta)$ as the collection of minimizers of $\int p(z, \theta) \|\nabla_\theta \log p(z, \theta) - s(z, \theta; \psi)\|^2 dz$ for the given θ . We further denote the empirical estimation function for the two terms, i.e.,

$$\int p(z, \theta) \left(\|s(z, \theta; \psi)\|^2 - 2\nabla_\theta \log p(z, \theta)^\top s(z, \theta; \psi) \right) dz,$$

as $\hat{J}_{n,k}(\psi; \theta)$ following the methods in Section 3.2 for any θ in the trajectory $\{\hat{\theta}_t\}_{t=1}^T$, where n and k are the number of samples we get at each iteration for $\hat{\theta}_t$ and perturbed policies.

Meaning of each assumption. Assumption A.1 is used to establish the asymptotic normality of our estimator under performativity. Additionally, we use the fact that the population loss Hessians $H_{\tilde{\theta}}(\theta) = \nabla_{\tilde{\theta}}^2 \mathbb{E}_{z \sim \mathcal{D}(\tilde{\theta})} \ell(z; \theta)$ are positive definite, which is guaranteed by the strong convexity of the loss function (Assumption 3.1). Assumption A.2 and A.3 are used in the analysis of score matching. Assumption A.2, based on the differentiation lemma [18], ensures the interchangeability of integration and differentiation. Assumption A.3 guarantees the consistency of the score matching estimator. Assumption A.4 and A.5 parallel to Assumption 3.1(a) and A.1, and are used to establish the consistency and asymptotic normality of our PPI estimator under performativity. Conditions such as local Lipschitz continuity and positive definiteness are standard for establishing asymptotic normality. Similar assumptions are also imposed in [1].

Assumption A.1 (Positive Definiteness & Regularity Conditions for the Estimator). We assume the following.

(a). The loss function satisfies the gradient covariance matrix is uniformly bounded below:

$$V_{\tilde{\theta}}(\theta) = \text{Cov}_{z \sim \mathcal{D}(\tilde{\theta})}(\nabla_\theta \ell(z; \theta)) \succeq cI,$$

for any $\tilde{\theta}$, where $c > 0$ is a constant.

(b). For any sample size n , assume the M-estimator $\hat{\theta}_t$ has a density function with respect to the Lebesgue measure, and its characteristic function is absolutely integrable.

Assumption A.2 (Regularity Condition for M). Assume that for $\forall i$:

(a). The function $z \mapsto p(z, \theta) \frac{\partial \log M(z, \theta; \psi)}{\partial \theta^{(i)}}$ is Lebesgue integrable.

(b). For almost every $z \in \mathcal{Z}$ (with respect to Lebesgue measure), the partial derivative

$$\frac{\partial}{\partial \theta^{(i)}} \left[p(z, \theta) \frac{\partial \log M(z, \theta; \psi)}{\partial \theta^{(i)}} \right]$$

exists.

(c). There exists a Lebesgue-integrable function $H(z)$ such that for almost every $z \in \mathcal{Z}$,

$$\left| \frac{\partial}{\partial \theta^{(i)}} \left[p(z, \theta) \frac{\partial \log M(z, \theta; \psi)}{\partial \theta^{(i)}} \right] \right| \leq H(z).$$

Assumption A.3 (Consistency of Optimizer). We let k grows along with n such that $n \rightarrow \infty$ leads to $k \rightarrow \infty$. We assume that the class $M(z, \theta; \psi)$ is rich enough that for all $\theta \in \Theta$, there exists $\psi^*(\theta)$ such that $M(z, \theta; \psi^*(\theta)) = p(z, \theta)$. Moreover, for the underlying trajectory $\{\theta_t\}_{t=1}^T$,

$$\lim_{n \rightarrow \infty} \arg \min_{\psi} \hat{J}_{n,k}(\psi; \hat{\theta}_t) \subseteq \Psi(\theta_t).$$

Assumption A.4 (Local Lipschitzness with f). Loss function $\ell(x, f(x); \theta)$ is locally Lipschitz: for each $\theta \in \Theta$, there exist a neighborhood $\Upsilon(\theta)$ of θ such that $\ell(x, f(x); \tilde{\theta})$ is $L^f(x)$ Lipschitz w.r.t $\tilde{\theta}$ for all $\tilde{\theta} \in \Upsilon(\theta)$ and $\mathbb{E}_{x \sim \mathcal{D}_X(\theta)} L^f(x) < \infty$.

Assumption A.5 (Positive Definiteness with f & Regularity Conditions for the PPI Estimator). We assume the following.

(a). Assume the loss function satisfies the the gradient covariance matrices are uniformly bounded below:

$$V_{\tilde{\theta}}(\theta) = \text{Cov}_{z \sim \mathcal{D}(\tilde{\theta})}(\nabla_{\theta} \ell(z; \theta)) \succeq cI, \quad V_{\tilde{\theta}}^f(\theta) = \text{Cov}_{x \sim \mathcal{D}_X(\tilde{\theta})}(\nabla_{\theta} \ell(x, f(x); \theta)) \succeq cI,$$

for any $\tilde{\theta}, \theta$, where $c > 0$ is a constant.

(b). For any sample size n , assume $\hat{\theta}_t^{\text{PPI}}$ has a density function with respect to the Lebesgue measure, and its characteristic function is absolutely integrable.

A.2 Details of Section 3: Theory of Inference under Performativity

We provide the omitted details in Section 3.

A.2.1 Consistency and Central Limit Theorem of $\hat{\theta}_t$

Let us denote:

$$\mathcal{L}_{\tilde{\theta}}(\theta) := \mathbb{E}_{z \sim \mathcal{D}(\tilde{\theta})} \ell(z; \theta), \quad \mathcal{L}_{\tilde{\theta}, n}(\theta) := \frac{1}{n} \sum_{i=1}^n \ell(z_i; \theta),$$

where the samples $z_i = (x_i, y_i) \sim \mathcal{D}(\tilde{\theta})$ are drawn from the distribution under $\tilde{\theta}$.

Proposition A.6 (Consistency of $\hat{\theta}_t$, Restatement of Proposition 3.3). *Under Assumption 3.1, if $\varepsilon < \frac{\gamma}{\beta}$, then for any given $T \geq 0$, we have that for all $t \in [T]$,*

$$\hat{\theta}_t \xrightarrow{P} \theta_t.$$

Proof. Let us denote $\hat{G}(\theta) := \text{argmin}_{\theta' \in \Theta} \frac{1}{n} \sum_{i=1}^n \ell(z_i; \theta')$ where the samples $z_i \sim \mathcal{D}(\theta)$ are drawn for some parameter θ along the dynamic trajectory $\theta_0 \rightarrow \hat{\theta}_1 \rightarrow \dots \hat{\theta}_t \rightarrow \dots$.

$$\begin{aligned} \|\theta_t - \hat{\theta}_t\| &= \|G(\theta_{t-1}) - \hat{G}(\hat{\theta}_{t-1})\| \\ &\leq \|G(\hat{\theta}_{t-1}) - \hat{G}(\hat{\theta}_{t-1})\| + \|G(\theta_{t-1}) - G(\hat{\theta}_{t-1})\| \\ &\leq \|G(\hat{\theta}_{t-1}) - \hat{G}(\hat{\theta}_{t-1})\| + \varepsilon \frac{\beta}{\gamma} \|\theta_{t-1} - \hat{\theta}_{t-1}\|, \end{aligned}$$

where the last inequality follows from the results derived by [25], under Assumption 3.1, we have $\|G(\theta) - G(\theta')\| \leq \frac{\varepsilon \beta}{\gamma} \|\theta - \theta'\|$.

Notice that $\mathbb{E}(\mathcal{L}_{\hat{\theta}_{t-1}, n}(\theta)) = \mathcal{L}_{\hat{\theta}_{t-1}}(\theta)$. By local Lipschitz condition, there exists $\varepsilon_0 > 0$ such that

$$\sup_{\theta: \|\theta - G(\hat{\theta}_{t-1})\| \leq \varepsilon_0} |\mathcal{L}_{\hat{\theta}_{t-1}, n}(\theta) - \mathcal{L}_{\hat{\theta}_{t-1}}(\theta)| \xrightarrow{P} 0.$$

Since ℓ is strongly convex for any θ , $G(\hat{\theta}_{t-1})$ is unique. Then we know that there exists δ such that $\mathcal{L}_{\hat{\theta}_{t-1}, n}(\theta) - \mathcal{L}_{\hat{\theta}_{t-1}}(G(\hat{\theta}_{t-1})) > \delta$ for all θ in $\{\theta \mid \|\theta - G(\hat{\theta}_{t-1})\| = \varepsilon_0\}$. Then it follows that:

$$\begin{aligned} &\inf_{\|\theta - G(\hat{\theta}_{t-1})\| = \varepsilon_0} \mathcal{L}_{\hat{\theta}_{t-1}, n}(\theta) - \mathcal{L}_{\hat{\theta}_{t-1}, n}(G(\hat{\theta}_{t-1})) \\ &= \inf_{\|\theta - G(\hat{\theta}_{t-1})\| = \varepsilon_0} \left((\mathcal{L}_{\hat{\theta}_{t-1}, n}(\theta) - \mathcal{L}_{\hat{\theta}_{t-1}}(\theta)) + (\mathcal{L}_{\hat{\theta}_{t-1}}(\theta) - \mathcal{L}_{\hat{\theta}_{t-1}}(G(\hat{\theta}_{t-1}))) \right. \\ &\quad \left. + (\mathcal{L}_{\hat{\theta}_{t-1}}(G(\hat{\theta}_{t-1})) - \mathcal{L}_{\hat{\theta}_{t-1}, n}(G(\hat{\theta}_{t-1}))) \right) \end{aligned}$$

$$\geq \delta - o_P(1).$$

Then we consider any fixed θ such that $\|\theta - G(\hat{\theta}_{t-1})\| \geq \varepsilon_0$ it follows that

$$\begin{aligned} \mathcal{L}_{\hat{\theta}_{t-1},n}(\theta) - \mathcal{L}_{\hat{\theta}_{t-1},n}(G(\hat{\theta}_{t-1})) &\geq \frac{\theta - G(\hat{\theta}_{t-1})}{\omega - G(\hat{\theta}_{t-1})} \left(\mathcal{L}_{\hat{\theta}_{t-1},n}(\omega) - \mathcal{L}_{\hat{\theta}_{t-1},n}(G(\hat{\theta}_{t-1})) \right) \\ &\geq \frac{\|\theta - G(\hat{\theta}_{t-1})\|}{\varepsilon_0} (\delta - o_P(1)) \geq \delta - o_P(1), \end{aligned}$$

where the first inequality holds for any ω by the convexity condition of $\mathcal{L}_{\hat{\theta}_{t-1},n}(\theta)$, and the second inequality holds as we take $\omega = \frac{\theta - G(\hat{\theta}_{t-1})}{\|\theta - G(\hat{\theta}_{t-1})\|} \varepsilon_0 + G(\hat{\theta}_{t-1})$ and using the above result. Thus no θ such that $\|\theta - G(\hat{\theta}_{t-1})\| = \varepsilon_0$ can be the minimizer of $\mathcal{L}_{\hat{\theta}_{t-1},n}(\theta)$. Then $\|G(\hat{\theta}_{t-1}) - \hat{G}(\hat{\theta}_{t-1})\| \xrightarrow{P} 0$.

We then have, for a given $T \geq 0$, we have that for all $t \in [T]$,

$$\|\hat{\theta}_t - \theta_t\| \leq \sum_{i=0}^t \left(\varepsilon \frac{\beta}{\gamma}\right)^{t-i} \|G(\hat{\theta}_i) - \hat{G}(\hat{\theta}_i)\| \xrightarrow{P} 0.$$

Thus, we conclude that $\hat{\theta}_t \xrightarrow{P} \theta_t$. ■

Theorem A.7 (Central Limit Theorem of $\hat{\theta}_t$, Restatement of Theorem 3.4). *Under Assumption 3.1 and A.1, if $\varepsilon < \frac{\gamma}{\beta}$, then for any given $T \geq 0$, we have that for all $t \in [T]$,*

$$\sqrt{n}(\hat{\theta}_t - \theta_t) \xrightarrow{D} \mathcal{N}(0, V_t)$$

with

$$V_t = \sum_{i=1}^t \left[\prod_{k=i}^{t-1} \nabla G(\theta_k) \right] \Sigma_{\theta_{i-1}}(\theta_i) \left[\prod_{k=i}^{t-1} \nabla G(\theta_k) \right]^\top.$$

In particular, $\nabla G(\theta_k) = -H_{\theta_k}(\theta_{k+1})^{-1}(\nabla_{\tilde{\theta}} \mathbb{E}_{z \sim \mathcal{D}(\theta_k)} \nabla_{\theta} \ell(z; \theta_{k+1}))$, where $\nabla_{\tilde{\theta}}$ is taking gradient for the parameter in $\mathcal{D}(\tilde{\theta})$, ∇_{θ} is taking gradient for the parameter in $\ell(z; \theta)$ and $\prod_{k=t}^{t-1} \nabla G(\theta_k) = I_d$.

Proof. Let $U_t := \sqrt{n}(\hat{\theta}_t - \theta_t)$ and denote $\tilde{\theta}_t = G(\hat{\theta}_{t-1})$. We make the following decomposition:

$$\hat{\theta}_t - \theta_t = \underbrace{(\tilde{\theta}_t - \theta_t)}_{(1)} + \underbrace{(\hat{\theta}_t - \tilde{\theta}_t)}_{(2)}.$$

Step 1: Conditional distribution of $U_t | U_{t-1}$.

For term (1), we have

$$\sqrt{n}(\tilde{\theta}_t - \theta_t) = \sqrt{n}(G(\hat{\theta}_{t-1}) - G(\theta_{t-1})).$$

For term (2), the empirical process analysis in [1] establishes that

$$\sqrt{n}(\hat{\theta}_t - \tilde{\theta}_t) | \hat{\theta}_{t-1} \xrightarrow{D} \mathcal{N}(0, \Sigma_{\hat{\theta}_{t-1}}(\tilde{\theta}_t)),$$

where the variance is given by

$$\Sigma_{\hat{\theta}_{t-1}}(\tilde{\theta}_t) = H_{\hat{\theta}_{t-1}}(\tilde{\theta}_t)^{-1} V_{\hat{\theta}_{t-1}}(\tilde{\theta}_t) H_{\hat{\theta}_{t-1}}(\tilde{\theta}_t)^{-1}.$$

Conditioning on $\hat{\theta}_{t-1}$ and considering the distribution $D(\hat{\theta}_{t-1})$, for any function h , we use the following shorthand notations:

$$\mathbb{E}_n h := \frac{1}{n} \sum_{i=1}^n h(x_i, y_i), \quad \mathbb{G}_n h := \sqrt{n}(\mathbb{E}_n h - \mathbb{E}_{(x,y) \sim \mathcal{D}(\hat{\theta}_{t-1})}[h(x, y)]).$$

Note that $\tilde{\theta}_t = G(\hat{\theta}_{t-1})$. Recall that

$$\mathcal{L}_{\tilde{\theta}}(\theta) := \mathbb{E}_{(x,y) \sim \mathcal{D}(\tilde{\theta})} \ell(x, y; \theta), \quad \mathcal{L}_{\tilde{\theta}, n} := \frac{1}{n} \sum_{i=1}^n \ell(x_i, y_i; \theta), \text{ where } (x_i, y_i) \sim \mathcal{D}(\tilde{\theta}).$$

Under the assumptions, Lemma 19.31 in [31] implies that for every sequence $h_n = O_P(1)$, we have

$$\mathbb{G}_n \left[\sqrt{n} \left(\ell(x, y; \tilde{\theta}_t + \frac{h_n}{\sqrt{n}}) - \ell(x, y; \tilde{\theta}_t) \right) - h_n^\top \nabla_{\theta} \ell(x, y; \tilde{\theta}_t) \right] \xrightarrow{P} 0.$$

Applying second-order Taylor expansion, we obtain that

$$\begin{aligned} n \mathbb{E}_n \left(\ell(x, y; \tilde{\theta}_t + \frac{h_n}{\sqrt{n}}) - \ell(x, y; \tilde{\theta}_t) \right) &= n \left(\mathcal{L}_{\hat{\theta}_{t-1}}(\tilde{\theta}_t + \frac{h_n}{\sqrt{n}}) - \mathcal{L}_{\hat{\theta}_{t-1}}(\tilde{\theta}_t) \right) \\ &\quad + h_n^\top \mathbb{G}_n \nabla_{\theta} \ell(x, y; \tilde{\theta}_t) + o_P(1) \\ &= \frac{1}{2} h_n^\top H_{\hat{\theta}_{t-1}}(\tilde{\theta}_t) h_n + h_n^\top \mathbb{G}_n \nabla_{\theta} \ell(x, y; \tilde{\theta}_t) + o_P(1). \end{aligned}$$

Set $h_n^* = \sqrt{n}(\hat{\theta}_t - \tilde{\theta}_t)$ and $h_n = -H_{\hat{\theta}_{t-1}}(\tilde{\theta}_t)^{-1} \mathbb{G}_n \nabla_{\theta} \ell(x, y; \tilde{\theta}_t)$, Corollary 5.53 in [31] implies they are $O_P(1)$.

Since $\hat{\theta}_t$ is the minimizer of $\mathcal{L}_{n, \hat{\theta}_{t-1}}$, the first term is smaller than the second term. We can rearrange the terms and obtain:

$$\frac{1}{2} (h_n^* - h_n)^\top H_{\hat{\theta}_{t-1}}(\tilde{\theta}_t) (h_n^* - h_n) = o_P(1),$$

which leads to $h_n^* - h_n = O_P(1)$. Then the above asymptotic normality result follows directly by applying the central limit theorem (CLT) to the following terms, conditioning on $\hat{\theta}_{t-1}$:

$$\begin{aligned} \sqrt{n}(\hat{\theta}_t - \tilde{\theta}_t) \mid \hat{\theta}_{t-1} &= -H_{\hat{\theta}_{t-1}}(\tilde{\theta}_t)^{-1} S + o_P(1), \\ S &= \sqrt{\frac{1}{n}} \sum_{i=1}^n \left(\nabla_{\theta} \ell(x_{t,i}, y_{t,i}; \tilde{\theta}_t) - \mathbb{E}_{(x,y) \sim \mathcal{D}(\hat{\theta}_{t-1})} [\nabla_{\theta} \ell(x, y; \tilde{\theta}_t)] \right). \end{aligned}$$

Note that, conditioning on $\hat{\theta}_{t-1}$, (1) is a constant. Therefore, (1) and (2) follow a joint Gaussian distribution. Consequently, given U_{t-1} , the conditional distribution of U_t is given by:

$$\begin{aligned} U_t \mid U_{t-1} &= \sqrt{n}(\hat{\theta}_t - \theta_t) \mid \hat{\theta}_{t-1} \\ &= \sqrt{n}(\tilde{\theta}_t - \theta_t) + \sqrt{n}(\hat{\theta}_t - \tilde{\theta}_t) \mid \hat{\theta}_{t-1} \\ &= \sqrt{n}(G(\hat{\theta}_{t-1}) - G(\theta_{t-1})) + \sqrt{n}(\hat{\theta}_t - \tilde{\theta}_t) \mid \hat{\theta}_{t-1} \\ &\xrightarrow{D} \mathcal{N} \left(\sqrt{n}(G(\hat{\theta}_{t-1}) - G(\theta_{t-1})), \Sigma_{\hat{\theta}_{t-1}}(\tilde{\theta}_t) \right). \\ &= \mathcal{N} \left(\sqrt{n} \left(G\left(\frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}\right) - G(\theta_{t-1}) \right), \Sigma_{\frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}} \left(G\left(\frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}\right) \right) \right). \end{aligned}$$

For later references, we denote $U_t \mid U_{t-1} \xrightarrow{D} \mathcal{N}(\mu(U_{t-1}), \Sigma(U_{t-1}))$.

Step 2: Marginal distribution of U_t . We calculate the characteristic function of U_t by induction. To begin with, we directly have

$$X_1 \xrightarrow{D} \mathcal{N}(0, V_1), \quad V_1 = \Sigma_{\theta_0}(\theta_1).$$

Now, assume that $U_{t-1} \xrightarrow{D} \mathcal{N}(0, V_{t-1})$, we derive the joint distribution of (U_t, U_{t-1}) and marginal distribution of U_t . Then we have, the characteristics functions ϕ and the probability density function p of distributions U_{t-1} and $U_t \mid U_{t-1}$ follow:

$$\phi_{U_{t-1}}(s) \rightarrow \phi_{\mathcal{N}(0, V_{t-1})}(s) = \exp\left(-\frac{1}{2} s^\top V_{t-1} s\right), \quad p_{U_{t-1}}(u) = \frac{1}{(2\pi)^d} \int e^{-iz^\top u} \phi_{U_{t-1}}(z) dz,$$

$$\phi_{U_t|U_{t-1}}(s) \rightarrow \phi_{\mathcal{N}(\mu(U_{t-1}), \Sigma(U_{t-1}))}(s) = \exp(is^T \mu(U_{t-1}) - \frac{1}{2}s^T \Sigma(U_{t-1})s).$$

Then we have

$$\begin{aligned} \phi_{U_t}(s) &= \mathbb{E}e^{is^T U_t} = \mathbb{E}(\mathbb{E}(e^{is^T U_t} | U_{t-1})) = E_{U_{t-1}} \phi_{U_t|U_{t-1}}(s | U_{t-1}) \\ &= \int \phi_{U_t|U_{t-1}}(s | u) p_{U_{t-1}}(u) du \\ &= \int \phi_{U_t|U_{t-1}}(s | u) \frac{1}{(2\pi)^d} \int e^{-iz^T u} \phi_{U_{t-1}}(z) dz du \\ &= \frac{1}{(2\pi)^d} \iint \phi_{U_t|U_{t-1}}(s | u) \phi_{U_{t-1}}(z) e^{-iz^T u} dz du \\ &= \frac{1}{(2\pi)^d} \iint \exp(is^T \mu(U_{t-1}) - \frac{1}{2}s^T \Sigma(U_{t-1})s) \exp(-\frac{1}{2}z^T V_{t-1}z) e^{-iz^T u} dz du \\ &= \frac{1}{(2\pi)^d} \int \exp(is^T \mu(U_{t-1}) - \frac{1}{2}s^T \Sigma(U_{t-1})s) \left(\int \exp(-\frac{1}{2}z^T V_{t-1}z - iz^T u) dz \right) du \\ &= \frac{1}{(2\pi)^d} \int \exp(is^T \mu(U_{t-1}) - \frac{1}{2}s^T \Sigma(U_{t-1})s) \\ &\quad \times \left(\int \exp(-\frac{1}{2}u^T V_{t-1}^{-1}u) \exp(-\frac{1}{2}(z - V_{t-1}^{-1}iu)^T V_{t-1}(z - V_{t-1}^{-1}iu)) dz \right) du \\ &= \frac{1}{(2\pi)^d} \int \exp(is^T \mu(U_{t-1}) - \frac{1}{2}s^T \Sigma(U_{t-1})s) ((2\pi)^{\frac{d}{2}} \frac{1}{\det |V_{t-1}|} \cdot \exp(-\frac{1}{2}u^T V_{t-1}^{-1}u)) du \\ &= \frac{1}{(2\pi)^{\frac{d}{2}} \det |V_{t-1}|} \int \exp(is^T \mu(U_{t-1}) - \frac{1}{2}s^T \Sigma(U_{t-1})s - \frac{1}{2}u^T V_{t-1}u) du. \end{aligned}$$

Apply dominant convergence theorem to $\lim_{n \rightarrow \infty} \phi_{U_t}(s)$, we have:

$$\begin{aligned} \lim_{n \rightarrow \infty} \phi_{U_t}(s) &= \lim_{n \rightarrow \infty} \frac{1}{(2\pi)^{\frac{d}{2}} \det |V_{t-1}|} \int \exp(is^T \mu(U_{t-1}) - \frac{1}{2}s^T \Sigma(U_{t-1})s - \frac{1}{2}u^T V_{t-1}u) du \\ &= \lim_{n \rightarrow \infty} \frac{1}{(2\pi)^{\frac{d}{2}} \det |V_{t-1}|} \int \exp(is^T \sqrt{n}(G(\frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}) - G(\theta_{t-1})) \\ &\quad - \frac{1}{2}s^T \Sigma_{\frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}}(G(\frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}))s - \frac{1}{2}u^T V_{t-1}u) du \\ &= \frac{1}{(2\pi)^{\frac{d}{2}} \det |V_{t-1}|} \int \lim_{n \rightarrow \infty} \exp(is^T \sqrt{n}(G(\frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}) - G(\theta_{t-1})) \\ &\quad - \frac{1}{2}s^T \Sigma_{\frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}}(G(\frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}))s - \frac{1}{2}u^T V_{t-1}u) du \\ &= \frac{1}{(2\pi)^{\frac{d}{2}} \det |V_{t-1}|} \int \exp(is^T \nabla G(\theta_{t-1})u - \frac{1}{2}s^T \Sigma_{\theta_{t-1}}(G(\theta_{t-1}))s - \frac{1}{2}u^T V_{t-1}u) du \\ &= \exp(-\frac{1}{2}s^T \nabla G(\theta_{t-1})V_{t-1} \nabla G(\theta_{t-1})^T s - \frac{1}{2}s^T \Sigma_{U_{t-1}}(\theta_t)s), \end{aligned}$$

which is the characteristic function of $\mathcal{N}(0, V_t)$, where $V_t = \nabla G(\theta_{t-1})V_{t-1} \nabla G(\theta_{t-1})^T + \Sigma_{\theta_{t-1}}(\theta_t)$. Here we use the fact that $\lim_{n \rightarrow \infty} \sqrt{n} \left(G(\frac{y}{\sqrt{n}} + \theta_{t-1}) - G(\theta_{t-1}) \right) = \nabla G(\theta_{t-1})y$, and the dominant convergence theorem holds as we have

$$\begin{aligned} &|\exp(is^T \sqrt{n}(G(\frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}) - G(\theta_{t-1})) \\ &\quad - \frac{1}{2}s^T \Sigma_{\frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}}(G(\frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}))s - \frac{1}{2}u^T V_{t-1}u)| \leq |\exp(-\frac{1}{2}u^T V_{t-1}u)|. \end{aligned}$$

Thus we conclude by induction that

$$U_t \xrightarrow{D} \mathcal{N}(0, V_t),$$

$$V_t = \sum_{i=1}^t \left[\prod_{k=i}^{t-1} \nabla G(\theta_k) \right] \Sigma_{\theta_{i-1}}(\theta_i) \left[\prod_{k=i}^{t-1} \nabla G(\theta_k) \right]^\top.$$

And by the theorem of implicit function, we can calculate the gradient of G as follows

$$\begin{aligned} \nabla G(\theta) &= - \left[\left(\nabla_{\psi}^2 \mathbb{E}_{(x,y) \sim \mathcal{D}(\theta)} \ell(x, y; \psi) \right) |_{\psi=G(\theta)} \right]^{-1} \left(\nabla_{\psi} \nabla_{\tilde{\theta}} \mathbb{E}_{(x,y) \sim \mathcal{D}(\theta)} \ell(x, y; \psi) \right) |_{\psi=G(\theta)}. \\ \nabla G(\theta_k) &= - \left[\mathbb{E}_{(x,y) \sim \mathcal{D}(\theta_k)} \nabla_{\tilde{\theta}}^2 \ell(x, y; \theta_{k+1}) \right]^{-1} \left(\nabla_{\tilde{\theta}} \mathbb{E}_{(x,y) \sim \mathcal{D}(\theta_k)} \nabla_{\theta} \ell(x, y; \theta_{k+1}) \right) \\ &= -H_{\theta_k}(\theta_{k+1})^{-1} \left(\nabla_{\tilde{\theta}} \mathbb{E}_{(x,y) \sim \mathcal{D}(\theta_k)} \nabla_{\theta} \ell(x, y; \theta_{k+1}) \right) \\ &= -H_{\theta_k}(\theta_{k+1})^{-1} \mathbb{E}_{z \sim \mathcal{D}(\theta_k)} [\nabla_{\theta} \ell(z, \theta_{k+1}) \nabla_{\theta} \log p(z, \theta_k)^\top]. \end{aligned}$$

■

A.2.2 Score Matching

In this part, we provide details about our score matching mechanism.

Given that

$$\nabla G(\theta_k) = -H_{\theta_k}(\theta_{k+1})^{-1} \mathbb{E}_{z \sim \mathcal{D}(\theta_k)} [\nabla_{\theta} \ell(z, \theta_{k+1}) \nabla_{\theta} \log p(z, \theta_k)^\top],$$

once we have a good estimation of $\nabla_{\theta} \log p(z, \theta_k)$ for all $z \in \mathcal{Z}$, $\nabla G(\theta_k)$ could be easily estimated by samples.

Recall that we use a model $M(z, \theta; \psi)$ parameterized by ψ to approximate $p(z, \theta)$. Inspired by the objective in [14], for any given θ (e.g., $\hat{\theta}_t$), we aim to optimize the following objective parameterized by ψ :

$$\begin{aligned} J(\theta; \psi) &= \int p(z, \theta) \|\nabla_{\theta} \log p(z, \theta) - s(z, \theta; \psi)\|^2 dz \\ &= \int p(z, \theta) \left(\|\nabla_{\theta} \log p(z, \theta)\|^2 + \|s(z, \theta; \psi)\|^2 - 2 \nabla_{\theta} \log p(z, \theta)^\top s(z, \theta; \psi) \right) dz \end{aligned}$$

where $s(z, \theta; \psi) = \nabla_{\theta} \log M(z, \theta; \psi)$.

As mentioned in the main context, the first term is unrelated to ψ ; the second term involves model M that is chosen by us, so we have the analytical expression of $s(z, \theta; \psi)$. Thus, our key task will be estimating the third term, which involves $\mathcal{K}(\theta; \psi) := \int p(z, \theta) \nabla_{\theta} \log p(z, \theta)^\top s(z, \theta; \psi) dz$.

Lemma A.8 (Restatement of Lemma 3.5). *Under Assumption A.2, we have*

$$\mathcal{K}(\theta; \psi) = \sum_{i=1}^d \left[\frac{\partial}{\partial \theta^{(i)}} \int p(z, \theta) \frac{\partial \log M(z, \theta; \psi)}{\partial \theta^{(i)}} dz - \int p(z, \theta) \frac{\partial^2 \log M(z, \theta; \psi)}{\partial \theta^{(i)2}} dz \right]$$

where $\theta^{(i)}$ is the i -th coordinate of θ .

Proof. Recall that θ is of d -dimension.

$$\begin{aligned} \int p(z, \theta) \nabla_{\theta} \log p(z, \theta)^\top s(z, \theta; \psi) dz &= \sum_{i=1}^d \int p(z, \theta) \frac{\partial \log p(z, \theta)}{\partial \theta^{(i)}} \cdot \frac{\partial \log M(z, \theta; \psi)}{\partial \theta^{(i)}} dz \\ &= \sum_{i=1}^d \int p(z, \theta) \frac{\partial \log p(z, \theta)}{\partial \theta^{(i)}} \cdot \frac{\partial \log M(z, \theta; \psi)}{\partial \theta^{(i)}} dz \\ &= \sum_{i=1}^d \int \frac{\partial p(z, \theta)}{\partial \theta^{(i)}} \cdot \frac{\partial \log M(z, \theta; \psi)}{\partial \theta^{(i)}} dz. \end{aligned}$$

Then, we study $\int \frac{\partial p(z, \theta)}{\partial \theta^{(i)}} \cdot \frac{\partial \log M(z, \theta; \psi)}{\partial \theta^{(i)}} dz$. Under Assumption A.2, the integral and differentiation of the following equation is exchangeable, i.e.,

$$\frac{\partial}{\partial \theta^{(i)}} \int p(z, \theta) \frac{\partial M(z, \theta; \psi)}{\partial \theta^{(i)}} dz = \int \frac{p(z, \theta)}{\partial \theta^{(i)}} \frac{\partial M(z, \theta; \psi)}{\partial \theta^{(i)}} dz.$$

According to integral by parts, we have

$$\int \frac{\partial p(z, \theta)}{\partial \theta^{(i)}} \cdot \frac{\partial \log M(z, \theta; \psi)}{\partial \theta^{(i)}} dz = \frac{\partial}{\partial \theta^{(i)}} \int p(z, \theta) \frac{\partial M(z, \theta; \psi)}{\partial \theta^{(i)}} dz - \int p(z, \theta) \frac{\partial^2 \log M(z, \theta; \psi)}{\partial \theta^{(i)2}} dz.$$

Thus, our proof is completed. $\blacksquare$

The rest of the estimation process via policy perturbation is provided in the main context in Section 3.

The other part omitted in Section 3 is the details about Eq. 2 that

$$\widehat{V}_t^{-1/2} \sqrt{n}(\widehat{\theta}_t - \theta_t) \xrightarrow{D} \mathcal{N}(0, I_d).$$

Here $\widehat{V}_t$ denotes the sample-based estimator of the variance, obtained by plugging in the empirical Hessian and empirical covariance matrices:

$$\widehat{H}_{\widehat{\theta}_{t-1}}(\widehat{\theta}_t) = \widehat{\mathbb{E}}_{z \sim \mathcal{D}(\widehat{\theta}_{t-1})} \nabla_{\theta}^2 \ell(z; \widehat{\theta}_t), \quad \widehat{\text{Cov}}_{z \sim \mathcal{D}(\widehat{\theta}_{t-1})}(\nabla_{\theta} \ell(z; \widehat{\theta}_t)),$$

as well as the estimator for $\nabla G(\widehat{\theta}_{t-1})$:

$$-\widehat{H}_{\widehat{\theta}_{t-1}}(\widehat{\theta}_t)^{-1} \widehat{\mathbb{E}}_{z \sim \mathcal{D}(\widehat{\theta}_{t-1})} [\nabla_{\theta} \ell(z, \widehat{\theta}_t) \nabla_{\theta} \log M(z, \widehat{\theta}_{t-1}, \widehat{\psi})^{\top}],$$

where $\widehat{\psi}$ is obtained by minimizing $\widehat{J}_{n,k}$.

Eq. 2 is a direct result following Slutsky's theorem. Assumption A.3 makes sure the empirical optimizer set can converge to the population optimizer set. Then other parts such as estimation of the Hessian matrix etc. could all be directly obtained by standard law of large numbers. Thus, we can directly use Slutsky's theorem to obtain Eq. 2.

A.3 Policy Perturbation

In this part, we prove the validity of $\widehat{g}_k$ as an estimator of $\nabla G(\theta_k)$. Recall that we have $s(z, \theta; \psi) = \nabla_{\theta} \log M(z, \theta; \psi)$ and we further denote $s_i = \frac{\partial M(z, \theta; \psi)}{\partial \theta^{(i)}}$, and use $\mathbb{E}_{\theta}$, $\widehat{\mathbb{E}}_{\theta, n}$ for the expectation $\mathbb{E}_{z \sim D(\theta)}$ and empirical expectation $\mathbb{E}_{z \sim \widehat{D}_n(\theta)}$ respectively.

Firstly, we define the following function families:

$$\begin{aligned} \mathcal{F}_{1, \theta} &:= \{s(\cdot, \theta; \psi) : \psi \in \Psi\}, \\ \mathcal{F}_{2, \theta}^{(i)} &:= \left\{ \frac{\partial}{\partial \theta^{(i)}} s_i(\cdot, \theta; \psi) : \psi \in \Psi \right\}, \\ \mathcal{F}_{3, \theta}^{(i)} &:= \{s_i(\cdot, \theta; \psi) : \psi \in \Psi\}. \end{aligned}$$

Further, we set

$$\begin{aligned} \widehat{J}_{n,k}(\theta; \psi) &:= \widehat{\mathbb{E}}_{\theta, n} [\|s(z, \theta; \psi)\|^2] + \mathbb{E}_{\theta} [\|\nabla_{\theta} \log p(z, \theta)\|^2] + 2 \sum_{i=1}^d [\widehat{\mathbb{E}}_{\theta, n} [\frac{\partial}{\partial \theta^{(i)}} s_i(z, \theta; \psi)]] \\ &\quad - 2 \sum_{i=1}^d \frac{1}{\eta} \left(\widehat{\mathbb{E}}_{\theta + \eta e^{(i)}, k} [s_i(z, \theta + \eta e^{(i)}; \psi)] - \widehat{\mathbb{E}}_{\theta, n} [s_i(z, \theta; \psi)] \right), \\ J_n(\theta; \psi) &:= \mathbb{E}_{\theta} [\|s(z, \theta; \psi)\|^2] + \mathbb{E}_{\theta} [\|\nabla_{\theta} \log p(z, \theta)\|^2] + 2 \sum_{i=1}^d [\mathbb{E}_{\theta} [\frac{\partial}{\partial \theta^{(i)}} s_i(z, \theta; \psi)]] \\ &\quad - 2 \sum_{i=1}^d \frac{1}{\eta} \left(\mathbb{E}_{\theta + \eta e^{(i)}} [s_i(z, \theta + \eta e^{(i)}; \psi)] - \mathbb{E}_{\theta} [s_i(z, \theta; \psi)] \right). \end{aligned}$$

Assumption A.9. We assume that the score function s , distribution map D and the corresponding function families $\{\mathcal{F}_{1,\theta}, \mathcal{F}_{2,\theta}^{(i)}, \mathcal{F}_{3,\theta}^{(i)} : i = 1, \dots, d\}$ has the following properties:

(a). There is a positive constant $C > 0$, such that for $\forall \theta \in \Theta; \psi \in \Psi$,

$$\left| \frac{\partial}{\partial \theta^{(i)^2}} \mathbb{E}_\theta[s_i(z, \theta; \psi)] \right| \leq C < \infty.$$

(b). (*Enveloping function*) For $\forall \theta \in \Theta$, there is a function $H_\theta(z)$, such that $\mathbb{E}_\theta[H_\theta(z)^2] < \infty$, and for any function $f \in \mathcal{F}_{1,\theta} \cup (\cup_{i=1}^d \mathcal{F}_{2,\theta}^{(i)}) \cup (\cup_{i=1}^d \mathcal{F}_{3,\theta}^{(i)})$, we have

$$|f(z)| \leq H_\theta(z)$$

(c). (*θ -uniform Donsker*) Let $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|)$ denote the covering number, that is the minimal number of $\|\cdot\|$ -balls of radius ε needed to cover the set $\mathcal{F}$, there exists $\rho(\varepsilon) > 0$, such that

$$\int_0^1 \sqrt{\rho(\varepsilon)} d\varepsilon < \infty,$$

and

$$\sup_{\theta \in \Theta} \sup_Q \log \mathcal{N}(\varepsilon \|H_\theta(z)\|_{Q,2}, \mathcal{F}_\theta, L_2(Q)) < \rho(\varepsilon),$$

where $\mathcal{F}_\theta \in \{\mathcal{F}_{1,\theta}, \mathcal{F}_{2,\theta}^{(i)}, \mathcal{F}_{3,\theta}^{(i)} : i = 1, \dots, d\}$ and Q is any distribution on $\mathcal{Z}$.

(d). (*Vanishing optimization error*) There exists $a_n = o(1)$, such that

$$\widehat{J}_{n,k}(\theta; \widehat{\psi}(\theta)) \leq \min_{\psi \in \Psi} \widehat{J}_{n,k}(\theta; \psi) + a_n.$$

(e). (*Richness of class*) There exists $\psi^*(\theta)$ for each θ , s.t.

$$s(z, \theta; \psi^*(\theta)) = \nabla_\theta \log p(z, \theta).$$

Remark A.10. Assumption A.9(c) holds with $\rho(\varepsilon) = C \log \frac{1}{\varepsilon}$, when $\mathcal{F}_\theta$ is a VC-subgraph class for all $\mathcal{F}_\theta \in \{\mathcal{F}_{1,\theta}, \mathcal{F}_{2,\theta}^{(i)}, \mathcal{F}_{3,\theta}^{(i)} : i = 1, \dots, d\}$.

Assumption A.11. Assume that the following conditions hold:

(a). (*Smoothness*) $H_{\tilde{\theta}}(\theta)$ is L -joint smooth in $(\tilde{\theta}, \theta)$ for some $L < \infty$.

(b). There exists $C > 0$, such that

$$\sup_{\tilde{\theta}} \mathbb{E}_{\tilde{\theta}}[\nabla_\theta \ell(z; \theta_{PS})] \leq C < \infty$$

Lemma A.12. Under Assumption A.9(a), we have

$$\sup_{\theta; \psi} |J(\theta; \psi) - J_n(\theta; \psi)| \leq 2dC\eta.$$

Proof. By mean-value theorem, there exists $\eta^{(i)} \in [0, \eta]$, $1 \leq i \leq d$, such that

$$\begin{aligned} J(\theta; \psi) - J_n(\theta; \psi) &= 2 \sum_{i=1}^d \frac{1}{\eta} \left(\mathbb{E}_{\theta + \eta e^{(i)}}[s_i(z, \theta + \eta e^{(i)}; \psi)] - \mathbb{E}_\theta[s_i(z, \theta; \psi)] \right) - 2 \sum_{i=1}^d \left[\frac{\partial}{\partial \theta^{(i)}} \mathbb{E}_\theta[s_i(z, \theta; \psi)] \right] \\ &= 2 \sum_{i=1}^d \left[\frac{\partial}{\partial \theta^{(i)}} \mathbb{E}_{\theta + \eta^{(i)} e^{(i)}}[s_i(z, \theta + \eta^{(i)} e^{(i)}; \psi)] \right] - 2 \sum_{i=1}^d \left[\frac{\partial}{\partial \theta^{(i)}} \mathbb{E}_\theta[s_i(z, \theta; \psi)] \right] \\ &\leq 2C \left(\sum_{i=1}^d \eta^{(i)} \right) \\ &\leq 2dC\eta, \end{aligned}$$

the first inequality follows from a direct application of mean-value theorem. ■

Theorem A.13. Let $\Theta_{d,t} = \{\hat{\theta}_j, \hat{\theta}_j + \eta e^{(i)} : i = 1, \dots, d, j = 1, \dots, T\}$. Under Assumption A.9, for $\forall \theta \in \Theta_{d,t}$, the following inequality holds

$$\mathbb{E}_\theta [\|\nabla \log p(z, \theta) - s(z, \theta; \psi(\theta))\|^2 | \hat{\psi}(\theta)] = O_p\left(\frac{1}{\sqrt{n}} + \frac{1}{\eta \sqrt{\min(n, k)}} + \eta + a_n\right).$$

Proof. Fix $\theta \in \Theta$, by Dudley's uniform entropy bound, c.f. Corollary 19.35 in [31], we have

$$\begin{aligned} \mathbb{E}_\theta \left[\sup_{\psi} \left| \hat{E}_{\theta,n} [\|s(z, \theta; \psi)\|^2] - E_\theta [\|s(z, \theta; \psi)\|^2] \right| \right] &\lesssim \frac{1}{\sqrt{n}}, \\ \mathbb{E}_\theta \left[\sup_{\psi} \left| \hat{E}_{\theta,n} \left[\frac{\partial}{\partial \theta^{(i)}} s_i(z, \theta; \psi) \right] - E_\theta \left[\frac{\partial}{\partial \theta^{(i)}} s_i(z, \theta; \psi) \right] \right| \right] &\lesssim \frac{1}{\sqrt{n}}, \\ \mathbb{E}_\theta \left[\sup_{\psi} \left| \hat{E}_{\theta,k} [s_i(z, \theta; \psi)\|^2] - E_\theta [\|s_i(z, \theta; \psi)\|^2] \right| \right] &\lesssim \frac{1}{\sqrt{k}}. \end{aligned}$$

Since $d, T = O(1)$, from the above results, the inequality below holds

$$\sup_{\psi \in \Psi} \sup_{\theta \in \Theta_{d,T}} |\hat{J}_{n,k}(\theta; \psi) - J_n(\theta; \psi)| = O_p\left(\frac{1}{\sqrt{n}} + \frac{1}{\eta \sqrt{\min(n, k)}}\right),$$

by Lemma A.12, we know

$$\sup_{\psi \in \Psi} \sup_{\theta \in \Theta_{d,T}} |\hat{J}_{n,k}(\theta; \psi) - J(\theta; \psi)| = O_p\left(\frac{1}{\sqrt{n}} + \frac{1}{\eta \sqrt{\min(n, k)}} + \eta\right). \quad (4)$$

From Assumption A.9(d), we know that for any $\theta \in \Theta_{d,T}$,

$$\hat{J}_{n,k}(\theta; \hat{\psi}(\theta)) \leq \min_{\psi \in \Psi} \hat{J}_{n,k}(\theta; \psi) + a_n,$$

Using Assumption A.9(e), we know that $\min_{\psi \in \Psi} J(\theta; \psi) = 0$, and assume there exists $\psi^*(\theta) \in \arg \min_{\psi \in \Psi} J(\theta; \psi)$.

Take $\psi = \psi^*(\theta)$ and $\hat{\psi}(\theta)$ respectively in inequality (4), we have

$$\begin{aligned} \hat{J}_{n,k}(\theta; \psi^*(\theta)) &= O_p\left(\frac{1}{\sqrt{n}} + \frac{1}{\eta \sqrt{\min(n, k)}} + \eta\right), \\ |J(\theta; \hat{\psi}(\theta)) - \hat{J}_{n,k}(\theta; \hat{\psi}(\theta))| &= O_p\left(\frac{1}{\sqrt{n}} + \frac{1}{\eta \sqrt{\min(n, k)}} + \eta\right). \end{aligned}$$

Since $\hat{J}_{n,k}(\theta; \hat{\psi}(\theta)) \leq \hat{J}_{n,k}(\theta; \psi^*(\theta)) + a_n$, we have

$$\begin{aligned} J(\theta; \hat{\psi}(\theta)) &= J(\theta; \hat{\psi}(\theta)) - \hat{J}_{n,k}(\theta; \hat{\psi}(\theta)) + \hat{J}_{n,k}(\theta; \hat{\psi}(\theta)) - \hat{J}_{n,k}(\theta; \psi^*(\theta)) + \hat{J}_{n,k}(\theta; \psi^*(\theta)) \\ &= O_p\left(\frac{1}{\sqrt{n}} + \frac{1}{\eta \sqrt{\min(n, k)}} + \eta + a_n\right). \end{aligned}$$

By definition of $J(\theta; \psi)$, we have proved

$$\mathbb{E}_\theta [\|\nabla \log p(z, \theta) - s(z, \theta; \psi(\theta))\|^2 | \hat{\psi}(\theta)] = O_p\left(\frac{1}{\sqrt{n}} + \frac{1}{\eta \sqrt{\min(n, k)}} + \eta + a_n\right).$$

■

Recall that

$$\nabla G(\theta_k) = -H_{\theta_k}(\theta_{k+1})^{-1} \mathbb{E}_{\theta_k} [\nabla_\theta \ell(z; \theta_{k+1}) \nabla_\theta \log p(z, \theta_k)^\top],$$

and the estimator is defined by

$$\hat{g}_k := -H_{\hat{\theta}_k}(\hat{\theta}_{k+1})^{-1} \hat{\mathbb{E}}_{\hat{\theta}_{k,n}} [\nabla_\theta \ell(z; \hat{\theta}_{k+1}) s(z, \hat{\theta}_k; \hat{\psi}(\hat{\theta}_k))^\top].$$

Theorem A.14 (Restatement of Theorem 3.6). *Under Assumption 3.1 A.1, A.9 and A.11, we have*

$$\|\hat{g}_k - \nabla G(\theta_k)\|^2 = O_p\left(\frac{1}{\sqrt{n}} + \frac{1}{\eta\sqrt{\min(n, k)}} + \eta + a_n\right).$$

Proof. By Assumption 3.1,

$$\|\nabla_{\theta}\ell(z; \theta_{k+1}) - \nabla_{\theta}\ell(z; \hat{\theta}_{k+1})\| \leq \beta\|\theta_{k+1} - \hat{\theta}_{k+1}\|,$$

by Assumption 3.1, there exists $C_1 > 0$, such that

$$\|H_{\hat{\theta}_k}(\hat{\theta}_{k+1})^{-1}\| \leq C_1 < \infty.$$

The proof falls into five parts.

Let

$$\hat{g}_{k,1} := -H_{\hat{\theta}_k}(\hat{\theta}_{k+1})^{-1}\hat{\mathbb{E}}_{\hat{\theta}_{k,n}}[\nabla_{\theta}\ell(z; \theta_{k+1})s(z, \hat{\theta}_k; \hat{\psi}(\hat{\theta}_k))^{\top}].$$

Step 1: Convergence of $\|\hat{g}_k - \hat{g}_{k,1}\|$.

We have

$$\begin{aligned}\|\hat{g}_k - \hat{g}_{k,1}\| &\lesssim \|\hat{\theta}_{k+1} - \theta_{k+1}\| \|\hat{\mathbb{E}}_{\hat{\theta}_{k,n}}[\|s(z, \hat{\theta}_k; \hat{\psi}(\hat{\theta}_k))\|]\| \\ &= \frac{1}{n} \|\hat{\theta}_{k+1} - \theta_{k+1}\| \sum_{l=1}^n \|s(z_{l,k}, \hat{\theta}_k; \hat{\psi}(\hat{\theta}_k))\|.\end{aligned}$$

From Assumption A.9(b), there exists $C_2 > 0$, such that

$$\mathbb{E}_{\hat{\theta}_k}[\|s(z, \hat{\theta}_k; \hat{\psi}(\hat{\theta}_k))\|^2] \leq C_2 < \infty,$$

by the law of large numbers, we know

$$\|\hat{g}_k - \hat{g}_{k,1}\| = O_p(\|\hat{\theta}_{k+1} - \theta_{k+1}\|),$$

thus combine the above result with Theorem 3.4, we have

$$\|\hat{g}_k - \hat{g}_{k,1}\| = O_p\left(\frac{1}{\sqrt{n}}\right).$$

Let

$$\hat{g}_{k,2} := -H_{\hat{\theta}_k}(\hat{\theta}_{k+1})^{-1}\hat{\mathbb{E}}_{\hat{\theta}_k}[\nabla_{\theta}\ell(z; \theta_{k+1})s(z, \hat{\theta}_k; \hat{\psi}(\hat{\theta}_k))^{\top}].$$

Step 2: Convergence of $\|\hat{g}_{k,1} - \hat{g}_{k,2}\|$.

By Cauchy-Schwarz inequality, we know

$$\mathbb{E}_{\hat{\theta}_k}[\|\nabla_{\theta}\ell(z; \theta_{k+1})s(z, \hat{\theta}_k; \hat{\psi}(\hat{\theta}_k))^{\top}\|^2] \leq \mathbb{E}_{\hat{\theta}_k}[\|\nabla_{\theta}\ell(z; \theta_{k+1})\|^2] \mathbb{E}_{\hat{\theta}_k}[\|s(z, \hat{\theta}_k; \hat{\psi}(\hat{\theta}_k))\|^2].$$

From Assumption A.9(b), we have

$$\mathbb{E}_{\hat{\theta}_k}[\|s(z, \hat{\theta}_k; \hat{\psi}(\hat{\theta}_k))\|^2] \leq \mathbb{E}_{\hat{\theta}_k}[H_{\hat{\theta}_k}^2(z)] \leq C < \infty,$$

where $H_{\hat{\theta}_k}(z)$ is the enveloping function for $\hat{\theta}_k$.

By Assumption 3.1,

$$\|\nabla_{\theta}\ell(z; \theta_{k+1}) - \nabla_{\theta}\ell(z; \theta_{PS})\| \leq \beta\|\theta_{k+1} - \theta_{PS}\|,$$

since

$$\|\theta_{k+1} - \theta_{PS}\| \leq \left(\frac{\varepsilon\beta}{\gamma}\right)^{k+1}\|\theta_0 - \theta_{PS}\| \leq \|\theta_0 - \theta_{PS}\|,$$

thus by Assumption A.11(b),

$$\mathbb{E}_{\hat{\theta}_k}[\|\nabla_{\theta}\ell(z; \theta_{k+1})\|^2] \lesssim \mathbb{E}_{\hat{\theta}_k}[\|\nabla_{\theta}\ell(z; \theta_{PS})\|^2] + \|\theta_0 - \theta_{PS}\|^2 = O(1). \quad (5)$$

Hence we have

$$\mathbb{E}_{\hat{\theta}_k}[\|\nabla_{\theta}\ell(z; \theta_{k+1})s(z, \hat{\theta}_k; \hat{\psi}(\hat{\theta}_k))^{\top}\|^2] = O(1), \quad (6)$$

by Chebyshev inequality and Assumption 3.1, this lead to the following bound of $\|\hat{g}_{k,1} - \hat{g}_{k,2}\|$,

$$\begin{aligned} \|\hat{g}_{k,1} - \hat{g}_{k,2}\| &\lesssim \left\| \mathbb{E}_{\hat{\theta}_k, n} [\nabla_{\theta} \ell(z; \theta_{k+1}) s(z, \hat{\theta}_k; \hat{\psi}(\hat{\theta}_k))^{\top}] - \mathbb{E}_{\hat{\theta}_k} [\nabla_{\theta} \ell(z; \theta_{k+1}) s(z, \hat{\theta}_k; \hat{\psi}(\hat{\theta}_k))^{\top}] \right\| \\ &= O_p\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

Let

$$\hat{g}_{k,3} := -H_{\theta_k}(\theta_{k+1})^{-1} \mathbb{E}_{\hat{\theta}_k} [\nabla_{\theta} \ell(z; \theta_{k+1}) s(z, \hat{\theta}_k; \hat{\psi}(\hat{\theta}_k))^{\top}].$$

Step 3: Convergence of $\|\hat{g}_{k,2} - \hat{g}_{k,3}\|$. By Assumption A.11(a) and (6), we have

$$\begin{aligned} \|\hat{g}_{k,2} - \hat{g}_{k,3}\| &\lesssim \|H_{\theta_k}(\theta_{k+1})^{-1} - H_{\hat{\theta}_k}(\hat{\theta}_{k+1})^{-1}\| \\ &= \|H_{\theta_k}(\theta_{k+1})^{-1} (H_{\theta_k}(\theta_{k+1}) - H_{\hat{\theta}_k}(\hat{\theta}_{k+1})) H_{\hat{\theta}_k}(\hat{\theta}_{k+1})^{-1}\|, \end{aligned}$$

further, using Assumption 3.1 and A.11(a), we know

$$\|\hat{g}_{k,2} - \hat{g}_{k,3}\| \lesssim \|\hat{\theta}_k - \theta_k\| + \|\hat{\theta}_{k+1} - \theta_{k+1}\|,$$

thus by Theorem 3.4,

$$\|\hat{g}_{k,2} - \hat{g}_{k,3}\| = O_p\left(\frac{1}{\sqrt{n}}\right).$$

Let

$$\hat{g}_{k,4} := -H_{\theta_k}(\theta_{k+1})^{-1} \mathbb{E}_{\hat{\theta}_k} [\nabla_{\theta} \ell(z; \theta_{k+1}) \nabla_{\theta} \log p(z, \hat{\theta}_k)].$$

Step 4: Convergence of $\|\hat{g}_{k,3} - \hat{g}_{k,4}\|$.

By Assumption 3.1 and Assumption A.1, we have

$$\begin{aligned} \|\hat{g}_{k,3} - \hat{g}_{k,4}\|^2 &\lesssim \left\{ \mathbb{E}_{\hat{\theta}_k} \left[\|\nabla_{\theta} \ell(z; \theta_{k+1})\| \times \|s(z, \hat{\theta}_k; \hat{\psi}(\hat{\theta}_k)) - \nabla_{\theta} \log p(z, \hat{\theta}_k)\| \right] \right\}^2 \\ &\leq \mathbb{E}_{\hat{\theta}_k} [\|\nabla_{\theta} \ell(z; \theta_{k+1})\|^2] \\ &\quad \times \mathbb{E}_{\hat{\theta}_k} [\|s(z, \hat{\theta}_k; \hat{\psi}(\hat{\theta}_k)) - \nabla_{\theta} \log p(z, \hat{\theta}_k)\|^2] \quad (\text{by Cauchy - Schwarz inequality}) \\ &\lesssim \mathbb{E}_{\hat{\theta}_k} [\|s(z, \hat{\theta}_k; \hat{\psi}(\hat{\theta}_k)) - \nabla_{\theta} \log p(z, \hat{\theta}_k)\|^2] \quad (\text{by formula (5)}) \\ &= O_p\left(\frac{1}{\sqrt{n}} + \frac{1}{\eta \sqrt{\min(n, k)}} + \eta + a_n\right) \quad (\text{by Theorem A.13}). \end{aligned}$$

Step 5: Convergence of $\|\hat{g}_k - \nabla G(\theta_k)\|$.

Finally, recall that

$$\nabla G(\theta_k) = -H_{\theta_k}(\theta_{k+1})^{-1} \mathbb{E}_{\hat{\theta}_k} [\nabla_{\theta} \ell(z; \theta_{k+1}) \nabla_{\theta} \log p(z, \theta_k)],$$

by Assumption 3.1 and A.1,

$$\begin{aligned} \|\hat{g}_{k,4} - \nabla G(\theta_k)\| &\lesssim \mathbb{E}_{\hat{\theta}_k} [\|\nabla_{\theta} \ell(z; \theta_{k+1})\|] \times \|\hat{\theta}_k - \theta_k\| \\ &\lesssim \|\hat{\theta}_k - \theta_k\| \\ &= O_p\left(\frac{1}{\sqrt{n}}\right) \quad (\text{By Theorem 3.4}). \end{aligned}$$

Combining the above results, we thus have proved

$$\|\hat{g}_k - \nabla G(\theta_k)\|^2 = O_p\left(\frac{1}{\sqrt{n}} + \frac{1}{\eta \sqrt{\min(n, k)}} + \eta + a_n\right).$$

■

A.4 Details of Section 4: Theory of Prediction-Powered Inference under Performativity

Without loss of generality, we let $N_t = N$ and $n_t = n$ for all $t \in [T]$.

Let us denote

$$\mathcal{L}_{\tilde{\theta}}(\theta) := \mathbb{E}_{(x,y) \sim \mathcal{D}(\tilde{\theta})} \ell(x, y; \theta), \quad \mathcal{L}_{\tilde{\theta}}^{f, \lambda}(\theta) := \mathcal{L}_{\tilde{\theta}, n}(\theta) + \lambda \cdot (\tilde{\mathcal{L}}_{\tilde{\theta}, N}^f(\theta) - \mathcal{L}_{\tilde{\theta}, n}^f(\theta)),$$

where

$$\mathcal{L}_{\tilde{\theta}, n}(\theta) := \frac{1}{n} \sum_{i=1}^n \ell(x_i, y_i; \theta), \quad \mathcal{L}_{\tilde{\theta}, n}^f(\theta) := \frac{1}{n} \sum_{i=1}^n \ell((x_i, f(x_i)); \theta), \quad \tilde{\mathcal{L}}_{\tilde{\theta}, N}^f(\theta) := \frac{1}{N} \sum_{i=1}^N \ell((x_i^u, f(x_i^u)); \theta).$$

Here the samples $(x_i, y_i) \sim \mathcal{D}(\tilde{\theta})$ and $x_i^u \sim \mathcal{D}_{\mathcal{X}}(\tilde{\theta})$ are drawn from the distribution under $\tilde{\theta}$. Recall that we have defined $\Sigma_{\lambda, \tilde{\theta}}(\theta) = H_{\tilde{\theta}}(\theta)^{-1} \left(r V_{\lambda, \tilde{\theta}}^f(\theta) + V_{\lambda, \tilde{\theta}}(\theta) \right) H_{\tilde{\theta}}(\theta)^{-1}$ before Theorem 4.1 (in the following we sometimes omit r for simplicity).

Theorem A.15 (Consistency of $\hat{\theta}_t^{\text{PPI}}$). *Under Assumption 3.1 and A.4, if $\varepsilon < \frac{\gamma}{\beta}$, then for any given $T \geq 0$, we have that for all $t \in [T]$,*

$$\hat{\theta}_{t+1}^{\text{PPI}}(\lambda_t) \xrightarrow{P} \theta_{t+1}.$$

Proof. Let us denote $\hat{G}_{\lambda}^f(\theta) := \arg \min_{\theta' \in \Theta} \frac{\lambda}{N} \sum_{i=1}^N \ell(x_i^u, f(x_i^u); \theta') + \frac{1}{n} \sum_{i=1}^n (\ell(x_i, y_i; \theta') - \lambda \ell(x_i, f(x_i); \theta'))$, where the samples $(x_i, y_i) \sim \mathcal{D}(\theta)$ and $x_i^u \sim \mathcal{D}_{\mathcal{X}}(\theta)$ are drawn for some parameter θ along the dynamic trajectory $\theta_0 \rightarrow \hat{\theta}_1 \rightarrow \dots \hat{\theta}_t \rightarrow \dots$.

$$\begin{aligned} \|\theta_t - \hat{\theta}_t^{\text{PPI}}\| &= \|G(\theta_{t-1}) - \hat{G}_{\lambda_t}^f(\hat{\theta}_{t-1}^{\text{PPI}})\| \\ &\leq \|G(\hat{\theta}_{t-1}^{\text{PPI}}) - \hat{G}_{\lambda_t}^f(\hat{\theta}_{t-1}^{\text{PPI}})\| + \|G(\theta_{t-1}) - G(\hat{\theta}_{t-1}^{\text{PPI}})\| \\ &\leq \|G(\hat{\theta}_{t-1}^{\text{PPI}}) - \hat{G}_{\lambda_t}^f(\hat{\theta}_{t-1}^{\text{PPI}})\| + \varepsilon \frac{\beta}{\gamma} \|\theta_{t-1} - \hat{\theta}_{t-1}^{\text{PPI}}\|, \end{aligned}$$

where the last inequality follows from the results derived by [25], under Assumption 3.1, we have $\|G(\theta) - G(\theta')\| \leq \frac{\varepsilon \beta}{\gamma} \|\theta - \theta'\|$.

Notice that $\mathbb{E}(\mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}^{f, \lambda_t}(\theta)) = \mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}(\theta)$. By local Lipschitz condition, there exists $\varepsilon_0 > 0$ such that

$$\sup_{\theta: \|\theta - G(\hat{\theta}_{t-1}^{\text{PPI}})\| \leq \varepsilon_0} |\mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}^{f, \lambda_t}(\theta) - \mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}(\theta)| \xrightarrow{P} 0.$$

Since ℓ is strongly convex for any θ , $G(\hat{\theta}_{t-1}^{\text{PPI}})$ is unique. Then we know that there exists δ such that $\mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}^{f, \lambda_t}(\theta) - \mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}(\hat{G}_{\lambda_t}^f(\hat{\theta}_{t-1}^{\text{PPI}})) > \delta$ for all θ in $\{\theta \mid \|\theta - G(\hat{\theta}_{t-1}^{\text{PPI}})\| = \varepsilon_0\}$. Then it follows that:

$$\begin{aligned} &\inf_{\|\theta - G(\hat{\theta}_{t-1}^{\text{PPI}})\| = \varepsilon_0} \mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}^{f, \lambda_t}(\theta) - \mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}(\hat{G}_{\lambda_t}^f(\hat{\theta}_{t-1}^{\text{PPI}})) \\ &= \inf_{\|\theta - G(\hat{\theta}_{t-1}^{\text{PPI}})\| = \varepsilon_0} \left((\mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}^{f, \lambda_t}(\theta) - \mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}(\theta)) + (\mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}(\theta) - \mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}(G(\hat{\theta}_{t-1}^{\text{PPI}}))) \right. \\ &\quad \left. + (\mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}(G(\hat{\theta}_{t-1}^{\text{PPI}})) - \mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}^{f, \lambda_t}(G(\hat{\theta}_{t-1}^{\text{PPI}}))) \right) \\ &\geq \delta - o_P(1). \end{aligned}$$

Then we consider any fixed θ such that $\|\theta - G(\hat{\theta}_{t-1}^{\text{PPI}})\| \geq \varepsilon_0$ it follows that

$$\begin{aligned} \mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}^{f, \lambda_t}(\theta) - \mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}(\hat{G}_{\lambda_t}^f(\hat{\theta}_{t-1}^{\text{PPI}})) &\geq \frac{\theta - G(\hat{\theta}_{t-1}^{\text{PPI}})}{\omega - G(\hat{\theta}_{t-1}^{\text{PPI}})} \left(\mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}^{f, \lambda_t}(\omega) - \mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}^{f, \lambda_t}(G(\hat{\theta}_{t-1}^{\text{PPI}})) \right) \\ &\geq \frac{\|\theta - G(\hat{\theta}_{t-1}^{\text{PPI}})\|}{\varepsilon_0} (\delta - o_P(1)) \geq \delta - o_P(1), \end{aligned}$$

where the first inequality holds for any ω by the convexity condition of $\mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}^{f, \lambda_t}(\theta)$, and the second inequality holds as we take $\omega = \frac{\theta - G(\hat{\theta}_{t-1}^{\text{PPI}})}{\|\theta - G(\hat{\theta}_{t-1}^{\text{PPI}})\|} \varepsilon_0 + G(\hat{\theta}_{t-1}^{\text{PPI}})$ and using the above result. Thus no θ such that $\|\theta - G(\hat{\theta}_{t-1}^{\text{PPI}})\| = \varepsilon_0$ can be the minimizer of $\mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}^{f, \lambda_t}(\theta)$. Then $\|G(\hat{\theta}_{t-1}^{\text{PPI}}) - \hat{G}_{\lambda_t}^f(\hat{\theta}_{t-1}^{\text{PPI}})\| \xrightarrow{P} 0$.

We then have, for a given $T \geq 0$, we have that for all $t \in [T]$,

$$\|\hat{\theta}_t^{\text{PPI}} - \theta_t\| \leq \sum_{i=0}^t \left(\varepsilon \frac{\beta}{\gamma}\right)^{t-i} \|G(\hat{\theta}_i^{\text{PPI}}) - \hat{G}_{\lambda_i}^f(\hat{\theta}_i^{\text{PPI}})\| \xrightarrow{P} 0.$$

Thus, we conclude that $\hat{\theta}_t^{\text{PPI}} \xrightarrow{P} \theta_t$. ■

Theorem A.16 (Central Limit Theorem of $\hat{\theta}_t^{\text{PPI}}(\lambda_t)$, Restatement of Theorem 4.1). *Under Assumption 3.1, A.4, and A.5, if $\varepsilon < \frac{\gamma}{\beta}$ and $\frac{n}{N} \rightarrow r$ for some $r \geq 0$, then for any given $T \geq 0$, we have that for all $t \in [T]$,*

$$\sqrt{n}(\hat{\theta}_t^{\text{PPI}}(\lambda_t) - \theta_t) \xrightarrow{D} \mathcal{N}\left(0, V_t^{\text{PPI}}(\{\lambda_j, \theta_j\}_{j=1}^t; r)\right)$$

with

$$V_t^{\text{PPI}}(\{\lambda_j, \theta_j\}_{j=1}^t; r) = \sum_{i=1}^t \left[\prod_{k=i}^{t-1} \nabla G(\theta_k) \right] \Sigma_{\lambda_i, \theta_{i-1}}(\theta_i; r) \left[\prod_{k=i}^{t-1} \nabla G(\theta_k) \right]^\top.$$

Proof. Let us denote the variance terms by V_t^{PPI} for simplicity, while omitting explicit dependence on parameters in the notation. Let $U_t := \sqrt{n}(\hat{\theta}_t^{\text{PPI}} - \theta_t)$ and denote $\tilde{\theta}_t = G(\hat{\theta}_{t-1}^{\text{PPI}})$. We make the following decomposition:

$$\hat{\theta}_t^{\text{PPI}} - \theta_t = \underbrace{(\tilde{\theta}_t - \theta_t)}_{(1)} + \underbrace{(\hat{\theta}_t^{\text{PPI}} - \tilde{\theta}_t)}_{(2)}.$$

Step 1: Conditional distribution of $U_t | U_{t-1}$.

For term (1), we have

$$\sqrt{n}(\tilde{\theta}_t - \theta_t) = \sqrt{n}(G(\hat{\theta}_{t-1}^{\text{PPI}}) - G(\theta_{t-1})).$$

For term (2), the empirical process analysis in [1] establishes that

$$\sqrt{n}(\hat{\theta}_t^{\text{PPI}} - \tilde{\theta}_t) | \hat{\theta}_{t-1}^{\text{PPI}} \xrightarrow{D} \mathcal{N}(0, \Sigma_{\lambda_t, \hat{\theta}_{t-1}^{\text{PPI}}}(\tilde{\theta}_t; r)),$$

where the variance is given by

$$\Sigma_{\hat{\theta}_{t-1}^{\text{PPI}}}(\tilde{\theta}_t; r) = H_{\hat{\theta}_{t-1}^{\text{PPI}}}(\tilde{\theta}_t)^{-1} \left(r V_{\lambda_t, \hat{\theta}_{t-1}^{\text{PPI}}}^f(\tilde{\theta}_t) + V_{\lambda_t, \hat{\theta}_{t-1}^{\text{PPI}}}(\tilde{\theta}_t) \right) H_{\hat{\theta}_{t-1}^{\text{PPI}}}(\tilde{\theta}_t)^{-1}.$$

Conditioning on $\hat{\theta}_{t-1}^{\text{PPI}}$, for any function h , we use the following shorthand notations:

$$\begin{aligned} \mathbb{E}_n h &:= \frac{1}{n} \sum_{i=1}^n h(x_i, y_i), & \mathbb{G}_n h &:= \sqrt{n}(\mathbb{E}_n h - \mathbb{E}_{(x,y) \sim \mathcal{D}(\hat{\theta}_{t-1}^{\text{PPI}})}[h(x, y)]), \\ \hat{\mathbb{E}}_N^f h &:= \frac{1}{N} \sum_{i=1}^N h(x_i^u, f(x_i^u)), & \hat{\mathbb{G}}_N^f h &:= \sqrt{N}(\hat{\mathbb{E}}_N^f h - \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}(\hat{\theta}_{t-1}^{\text{PPI}})}[h(x, f(x))]), \\ \hat{\mathbb{E}}_n^f h &:= \frac{1}{n} \sum_{i=1}^n h(x_i, f(x_i)), & \hat{\mathbb{G}}_n^f h &:= \sqrt{n}(\hat{\mathbb{E}}_n^f h - \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}(\hat{\theta}_{t-1}^{\text{PPI}})}[h(x, f(x))]). \end{aligned}$$

Note that $\tilde{\theta}_t = G(\hat{\theta}_{t-1}^{\text{PPI}})$. Recall that

$$\mathcal{L}_{\tilde{\theta}}(\theta) := \mathbb{E}_{(x,y) \sim \mathcal{D}(\tilde{\theta})} \ell(x, y; \theta), \quad \mathcal{L}_{\tilde{\theta}}^{f, \lambda}(\theta) := \mathcal{L}_{\tilde{\theta}, n}(\theta) + \lambda \cdot (\tilde{\mathcal{L}}_{\tilde{\theta}, N}^f(\theta) - \mathcal{L}_{\tilde{\theta}, n}^f(\theta)).$$

Under the assumptions, Lemma 19.31 in [31] implies that for every sequence $h_n = O_P(1)$, we have

$$\begin{aligned}\mathbb{G}_n \left[\sqrt{n} \left(\ell(x, y; \tilde{\theta}_t + \frac{h_n}{\sqrt{n}}) - \ell(x, y; \tilde{\theta}_t) \right) - h_n^\top \nabla_\theta \ell(x, y; \tilde{\theta}_t) \right] &\xrightarrow{P} 0, \\ \widehat{\mathbb{G}}_N^f \left[\sqrt{n} \left(\ell(x, y; \tilde{\theta}_t + \frac{h_n}{\sqrt{n}}) - \ell(x, y; \tilde{\theta}_t) \right) - h_n^\top \nabla_\theta \ell(x, y; \tilde{\theta}_t) \right] &\xrightarrow{P} 0, \\ \widehat{\mathbb{G}}_n^f \left[\sqrt{n} \left(\ell(x, y; \tilde{\theta}_t + \frac{h_n}{\sqrt{n}}) - \ell(x, y; \tilde{\theta}_t) \right) - h_n^\top \nabla_\theta \ell(x, y; \tilde{\theta}_t) \right] &\xrightarrow{P} 0.\end{aligned}$$

Applying second-order Taylor expansion, we obtain that

$$\begin{aligned}n\mathbb{E}_n \left(\ell(x, y; \tilde{\theta}_t + \frac{h_n}{\sqrt{n}}) - \ell(x, y; \tilde{\theta}_t) \right) &= n \left(\mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}(\tilde{\theta}_t + \frac{h_n}{\sqrt{n}}) - \mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}(\tilde{\theta}_t) \right) + h_n^\top \mathbb{G}_n \nabla_\theta \ell(x, y; \tilde{\theta}_t) + o_p(1) \\ &= \frac{1}{2} h_n^\top H_{\hat{\theta}_{t-1}^{\text{PPI}}}(\tilde{\theta}_t) h_n + h_n^\top \mathbb{G}_n \nabla_\theta \ell(x, y; \tilde{\theta}_t) + o_p(1).\end{aligned}$$

Based on similar calculation of the previous two terms, we can obtain that:

$$\begin{aligned}&n \left(\mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}^{f, \lambda}(\tilde{\theta}_t + \frac{h_n}{\sqrt{n}}) - \mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}^{f, \lambda}(\tilde{\theta}_t) \right) \\ &= \frac{1}{2} h_n^\top H_{\hat{\theta}_{t-1}^{\text{PPI}}}(\tilde{\theta}_t) h_n + h_n^\top \left(\mathbb{G}_n + \lambda \sqrt{\frac{n}{N}} \widehat{\mathbb{G}}_N^f - \lambda \widehat{\mathbb{G}}_n^f \right) \nabla_\theta \ell(x, y; \tilde{\theta}_t) + o_p(1).\end{aligned}$$

By considering $h_n^* = \sqrt{n}(\hat{\theta}_t^{\text{PPI}} - \tilde{\theta}_t)$ and $h_n = -H_{\hat{\theta}_{t-1}^{\text{PPI}}}(\tilde{\theta}_t)^{-1} \left(\mathbb{G}_n + \lambda \sqrt{\frac{n}{N}} \widehat{\mathbb{G}}_N^f - \lambda \widehat{\mathbb{G}}_n^f \right) \nabla_\theta \ell(x, y; \tilde{\theta}_t)$, Corollary 5.53 in [31] implies they are $O_P(1)$ and we obtain that

$$\begin{aligned}n \left(\mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}^{f, \lambda}(\hat{\theta}_t^{\text{PPI}}) - \mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}^{f, \lambda}(\tilde{\theta}_t) \right) &= \frac{1}{2} h_n^{*\top} H_{\hat{\theta}_{t-1}^{\text{PPI}}}(\tilde{\theta}_t) h_n^* + h_n^{*\top} \left(\mathbb{G}_n + \lambda \sqrt{\frac{n}{N}} \widehat{\mathbb{G}}_N^f - \lambda \widehat{\mathbb{G}}_n^f \right) \nabla_\theta \ell(x, y; \tilde{\theta}_t) + o_p(1) \\ n \left(\mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}^{f, \lambda}(\tilde{\theta}_t + \frac{h_n}{\sqrt{n}}) - \mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}^{f, \lambda}(\tilde{\theta}_t) \right) &= -\frac{1}{2} h_n^\top H_{\hat{\theta}_{t-1}^{\text{PPI}}}(\tilde{\theta}_t) h_n + o_P(1).\end{aligned}$$

Since $\hat{\theta}_t^{\text{PPI}}$ is the minimizer of $\mathcal{L}_{\hat{\theta}_{t-1}^{\text{PPI}}}^{f, \lambda}$, the first term is smaller than the second term. We can rearrange the terms and obtain:

$$\frac{1}{2} (h_n^* - h_n)^\top H_{\hat{\theta}_{t-1}^{\text{PPI}}}(\tilde{\theta}_t) (h_n^* - h_n) = o_P(1),$$

which leads to $h_n^* - h_n = O_P(1)$. Then the above asymptotic normality result follows directly by applying the central limit theorem (CLT) to the following terms, conditioning on $\hat{\theta}_{t-1}^{\text{PPI}}$:

$$\begin{aligned}\sqrt{n}(\hat{\theta}_t^{\text{PPI}} - \tilde{\theta}_t) | \hat{\theta}_{t-1}^{\text{PPI}} &= -H_{\hat{\theta}_{t-1}^{\text{PPI}}}(\tilde{\theta}_t)^{-1} (S_1 + S_2) + o_P(1), \\ S_1 &= \lambda_t \sqrt{\frac{n}{N}} \sqrt{\frac{1}{N}} \sum_{i=1}^N \left(\nabla_\theta \ell(x_{t,i}^u, f(x_{t,i}^u); \tilde{\theta}_t) - \mathbb{E}_{x \sim \mathcal{D}_X(\hat{\theta}_{t-1}^{\text{PPI}})} \nabla_\theta \ell(x, f(x); \tilde{\theta}_t) \right), \\ S_2 &= \sqrt{\frac{1}{n}} \sum_{i=1}^n \left(\nabla_\theta \ell(x_{t,i}, y_{t,i}; \tilde{\theta}_t) - \lambda_t \nabla_\theta \ell(x_{t,i}, f(x_{t,i}); \tilde{\theta}_t) \right. \\ &\quad \left. - \mathbb{E}_{(x,y) \sim \mathcal{D}(\hat{\theta}_{t-1}^{\text{PPI}})} [\nabla_\theta \ell(x, y; \tilde{\theta}_t) - \lambda_t \nabla_\theta \ell(x, f(x); \tilde{\theta}_t)] \right).\end{aligned}$$

Note that, conditioning on $\hat{\theta}_{t-1}^{\text{PPI}}$, (1) is a constant. Therefore, (1) and (2) follow a joint Gaussian distribution. Consequently, given U_{t-1} , the conditional distribution of U_t is given by:

$$\begin{aligned}U_t | U_{t-1} &= \sqrt{n}(\hat{\theta}_t^{\text{PPI}} - \theta_t) | \hat{\theta}_{t-1}^{\text{PPI}} \\ &= \sqrt{n}(\tilde{\theta}_t - \theta_t) + \sqrt{n}(\hat{\theta}_t^{\text{PPI}} - \tilde{\theta}_t) | \hat{\theta}_{t-1}^{\text{PPI}} \\ &= \sqrt{n}(G(\hat{\theta}_{t-1}^{\text{PPI}}) - G(\theta_{t-1})) + \sqrt{n}(\hat{\theta}_t^{\text{PPI}} - \tilde{\theta}_t) | \hat{\theta}_{t-1}^{\text{PPI}}\end{aligned}$$

$$\begin{aligned} & \xrightarrow{D} \mathcal{N}\left(\sqrt{n}(G(\hat{\theta}_{t-1}^{\text{PPI}}) - G(\theta_{t-1})), \Sigma_{\lambda_t, \hat{\theta}_{t-1}^{\text{PPI}}}(\tilde{\theta}_t; r)\right). \\ & = \mathcal{N}\left(\sqrt{n}\left(G\left(\frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}\right) - G(\theta_{t-1})\right), \Sigma_{\lambda_t, \frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}}\left(G\left(\frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}\right); r\right)\right). \end{aligned}$$

For later references, we denote $U_t \mid U_{t-1} \xrightarrow{D} \mathcal{N}(\mu(U_{t-1}), \Sigma(U_{t-1}; r))$.

Step 2: Marginal distribution of U_t . We calculate the characteristic function of U_t by induction. To begin with, we directly have

$$X_1 \xrightarrow{D} \mathcal{N}(0, V_1^{\text{PPI}}), \quad V_1^{\text{PPI}} = \Sigma_{\lambda_0, \theta_0}(\theta_1; r).$$

Now, assume that $U_{t-1} \xrightarrow{D} \mathcal{N}(0, V_{t-1}^{\text{PPI}})$, we derive the joint distribution of (U_t, U_{t-1}) and marginal distribution of U_t . Then we have, the characteristics functions ϕ and the probability density function p of distributions U_{t-1} and $U_t \mid U_{t-1}$ follow:

$$\begin{aligned} \phi_{U_{t-1}}(s) & \rightarrow \phi_{\mathcal{N}(0, V_{t-1}^{\text{PPI}})}(s) = \exp\left(-\frac{1}{2}s^T V_{t-1}^{\text{PPI}} s\right), \quad p_{U_{t-1}}(u) = \frac{1}{(2\pi)^d} \int e^{-iz^T u} \phi_{U_{t-1}}(z) dz \\ \phi_{U_t \mid U_{t-1}}(s) & \rightarrow \phi_{\mathcal{N}(\mu(U_{t-1}), \Sigma(U_{t-1}; r))}(s) = \exp\left(is^T \mu(U_{t-1}) - \frac{1}{2}s^T \Sigma(U_{t-1}; r) s\right). \end{aligned}$$

Then according to the proof of vanilla CLT under performativity in Section A.2, we have:

$$\phi_{U_t}(s) = \frac{1}{(2\pi)^{\frac{d}{2}} \det |V_{t-1}^{\text{PPI}}|} \int \exp\left(is^T \mu(U_{t-1}) - \frac{1}{2}s^T \Sigma(U_{t-1}; r) s - \frac{1}{2}u^T V_{t-1}^{\text{PPI}} u\right) du.$$

Apply dominant convergence theorem to $\lim_{n \rightarrow \infty} \phi_{U_t}(s)$, we have:

$$\begin{aligned} \lim_{n \rightarrow \infty} \phi_{U_t}(s) & = \lim_{n \rightarrow \infty} \frac{1}{(2\pi)^{\frac{d}{2}} \det |V_{t-1}^{\text{PPI}}|} \int \exp\left(is^T \mu(U_{t-1}) - \frac{1}{2}s^T \Sigma(U_{t-1}; r) s - \frac{1}{2}u^T V_{t-1}^{\text{PPI}} u\right) du \\ & = \lim_{n \rightarrow \infty} \frac{1}{(2\pi)^{\frac{d}{2}} \det |V_{t-1}^{\text{PPI}}|} \int \exp\left(is^T \sqrt{n}\left(G\left(\frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}\right) - G(\theta_{t-1})\right) \right. \\ & \quad \left. - \frac{1}{2}s^T \Sigma_{\lambda_t, \frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}}\left(G\left(\frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}\right); r\right) s - \frac{1}{2}u^T V_{t-1}^{\text{PPI}} u\right) du \\ & = \frac{1}{(2\pi)^{\frac{d}{2}} \det |V_{t-1}^{\text{PPI}}|} \int \exp\left(is^T \sqrt{n}\left(G\left(\frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}\right) - G(\theta_{t-1})\right) \right. \\ & \quad \left. - \frac{1}{2}s^T \Sigma_{\lambda_t, \frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}}\left(G\left(\frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}\right); r\right) s - \frac{1}{2}u^T V_{t-1}^{\text{PPI}} u\right) du \\ & = \frac{1}{(2\pi)^{\frac{d}{2}} \det |V_{t-1}^{\text{PPI}}|} \int \exp\left(is^T \nabla G(\theta_{t-1}) u - \frac{1}{2}s^T \Sigma_{\lambda_t, \theta_{t-1}}(G(\theta_{t-1}); r) s - \frac{1}{2}u^T V_{t-1}^{\text{PPI}} u\right) du \\ & = \exp\left(-\frac{1}{2}s^T \nabla G(\theta_{t-1}) V_{t-1}^{\text{PPI}} \nabla G(\theta_{t-1})^T s - \frac{1}{2}s^T \Sigma_{\lambda_t, \theta_{t-1}}(\theta_t; r) s\right), \end{aligned}$$

which is the characteristic function of $\mathcal{N}(0, V_t^{\text{PPI}})$, where $V_t^{\text{PPI}} = \nabla G(\theta_{t-1}) V_{t-1}^{\text{PPI}} \nabla G(\theta_{t-1})^T + \Sigma_{\lambda_t, \theta_{t-1}}(\theta_t; r)$. Here we use the fact that $\lim_{n \rightarrow \infty} \sqrt{n}\left(G\left(\frac{y}{\sqrt{n}} + \theta_{t-1}\right) - G(\theta_{t-1})\right) = \nabla G(\theta_{t-1}) y$, and the dominant convergence theorem holds as we have

$$\begin{aligned} & \left| \exp\left(is^T \sqrt{n}\left(G\left(\frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}\right) - G(\theta_{t-1})\right) - \frac{1}{2}s^T \Sigma_{\lambda_t, \frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}}\left(G\left(\frac{U_{t-1}}{\sqrt{n}} + \theta_{t-1}\right); r\right) s - \frac{1}{2}u^T V_{t-1}^{\text{PPI}} u\right) \right| \\ & \leq \left| \exp\left(-\frac{1}{2}u^T V_{t-1}^{\text{PPI}} u\right) \right|. \end{aligned}$$

Thus we conclude by induction that

$$\begin{aligned} U_t & \xrightarrow{D} \mathcal{N}(0, V_t^{\text{PPI}}), \\ V_t^{\text{PPI}} & = \sum_{i=1}^t \left[\prod_{k=i}^{t-1} \nabla G(\theta_k) \right] \Sigma_{\lambda_{i-1}, \theta_{i-1}}(\theta_i; r) \left[\prod_{k=i}^{t-1} \nabla G(\theta_k) \right]^T. \end{aligned}$$

And we have:

$$\begin{aligned}
\nabla G(\theta_k) &= - \left[\mathbb{E}_{(x,y) \sim \mathcal{D}(\theta_k)} \nabla_{\theta}^2 \ell(x, y; \theta_{k+1}) \right]^{-1} (\nabla_{\tilde{\theta}} \mathbb{E}_{(x,y) \sim \mathcal{D}(\theta_k)} \nabla_{\theta} \ell(x, y; \theta_{k+1})) \\
&= -H_{\theta_k}(\theta_{k+1})^{-1} (\nabla_{\tilde{\theta}} \mathbb{E}_{(x,y) \sim \mathcal{D}(\theta_k)} \nabla_{\theta} \ell(x, y; \theta_{k+1})) \\
&= -H_{\theta_k}(\theta_{k+1})^{-1} \mathbb{E}_{z \sim \mathcal{D}(\theta_k)} [\nabla_{\theta} \ell(z, \theta_{k+1}) \nabla_{\theta} \log p(z, \theta_k)^{\top}].
\end{aligned}$$

■

B Experimental Details

B.1 Additional Experimental Details

As described in Section 5, we construct simulation studies on a performative linear regression problem, where data are sampled from $D(\theta)$ as

$$y = \alpha^{\top} x + \mu^{\top} \theta + \nu, \quad x \sim \mathcal{N}(\mu_x, \Sigma_x), \quad \nu \sim \mathcal{N}(0, \sigma_y^2).$$

At each time step t , the label y_t is updated with $\hat{\theta}_{t-1}$ via the above equation, and then $\hat{\theta}_t$ is obtained by empirical repeated risk minimization with the updated data $z_t = (x_t, y_t)$. The objective of this task is to provide inference on an unbiased $\hat{\theta}_t$ with low variance, that is, the ground-truth θ_t is covered by the confidence region of $\hat{\theta}_t$ with high probability, and the width of this confidence region is small.

Given a set of labeled data, we can obtain the underlying θ_t as

$$\theta_t = (\Sigma_x + \mu_x \mu_x^{\top} + \gamma I_d)^{-1} (\mu_x \mu^{\top} \theta_{t-1} + (\Sigma_x + \mu_x \mu_x^{\top}) \alpha). \quad (7)$$

To compute the coverage and width of a confidence region $\hat{\mathcal{R}}_t(n, \delta)$ for θ_t , we run 1000 independent trials. For each trial j , we sample $\hat{\theta}_{t,j}$ together with its estimated variance $\hat{V}_{t,j}$, and construct two-sided normal intervals for each coordinate $i = 1, \dots, d$:

$$\left[\hat{\theta}_{t,j}^{(i)} \pm q_{1-\frac{\delta}{2d}} \sqrt{\hat{V}_{t,j}^{(i)} / n} \right], \quad q_{1-\frac{\delta}{2d}} = \Phi^{-1}(1 - \frac{\delta}{2d}),$$

where $d = 2$ is the parameter dimension, $\delta = 0.1$ the significance level, n the data size, and Φ^{-1} the standard normal quantile. The interval width of each trial is averaged over d coordinate intervals, and we count this trial as covered if the ground-truth θ_t lies inside *all* d coordinate intervals simultaneously. Finally, we report the average width and coverage rate over all trials.

Similarly, to compute the coverage for performative stable point θ_{PS} , we can obtain the close-form θ_{PS} for this task as follows:

$$\theta_{\text{PS}} = (\Sigma_x + \mu_x \mu_x^{\top} - \mu_x \mu^{\top} + \gamma I_d)^{-1} (\Sigma_x + \mu_x \mu_x^{\top}). \quad (8)$$

As defined in Corollary 3.7, the confidence region for θ_{PS} is constructed with

$$\hat{\mathcal{R}}_t(n, \delta) + \mathcal{B}\left(0, 2B\left(\frac{\varepsilon\beta}{\gamma}\right)^t\right),$$

where $\varepsilon = \|\mu\|_2$, $B = \|\theta_0 - \theta_{\text{PS}}\|_2$, and

$$\beta = \max \left\{ \max_{x \in \mathcal{X}} \{\|x\|_2^2 + \gamma\}, \max_{(x,y) \in (\mathcal{X}, \mathcal{Y}), \theta \in \Theta} \{\sqrt{(x^{\top} \theta - y + \|x\|_2 \|\theta\|_2)^2 + \|x\|_2^2}\} \right\}.$$

Here we take $\mathcal{X} = \{x : \|x\|_2^2 \leq 20\}$. Note that the closed form expressions for the update and the performatively stable point in Eq. 7 and Eq. 8 hold for any distribution of x with mean μ_x and variance Σ_x , and ν with mean 0 and variance σ_y^2 . For easier calculation for the smoothness parameter, we truncate the normal distribution of (x, y) such that $\|x\|_2^2 \leq 20$. The mean and variance of the resulting truncated distribution can be well approximated by those of the original normal distribution due to the concentration of Gaussian.

We run our experiments on NVIDIA GPUs A100 in a single-GPU setup.

B.2 Additional Experimental Results

Ablation study on effects of γ . In Figure A1, we compare confidence-region performance under regularization strengths $\gamma = 1$ and $\gamma = 3$. Together with results of $\gamma = 2$ in Figure 1, we can find that as γ increases, the gap between the coverage for θ_t (solid curve) and the bias-adjusted coverage for θ_{PS} (dashed curve) vanishes more quickly across iterations. For example, at $t = 3$, the dashed and solid curves are tightly closed for $\gamma = 3$, while a substantial gap remains for $\gamma = 1$. This phenomenon derives that the larger γ yields a more strongly convex loss, which both accelerates convergence of the estimate $\hat{\theta}_t$ to its stable point and reduces the performative bias $\|\theta_t - \theta_{PS}\|$. Consequently, the bias-aware intervals converge for θ_{PS} to the original ones for θ_t in fewer iterations when γ is larger.

Ablation study on effects of ε . In Figure A2, we compare confidence-region performance under sensitivity $\varepsilon \approx 0.003$ and $\varepsilon \approx 0.03$. We can find that as ε increases, the gap between the coverage for θ_t (solid curve) and the bias-adjusted coverage for θ_{PS} (dashed curve) vanishes more slowly across iterations. For example, for $\varepsilon \approx 0.003$, the dashed curves tightly upper-bound the solid curves at $t = 3$, whereas for $\varepsilon \approx 0.03$, a noticeable gap persists even at $t = 5$. This behavior is because a higher ε amplifies the performative shift (the dependence of the label distribution on θ), which increases the performative bias. That is, stronger sensitivity requires more iterations for $\hat{\theta}_t$ to approach its stable point, slowing down convergence of the two confidence regions.

Ablation study on effects of σ_y^2 . In Figure A4, we compare confidence-region performance under noise level $\sigma_y^2 = 0.1$ and $\sigma_y^2 = 0.4$. We observe that across all settings, PPI with our greedy-selected $\hat{\lambda}$ is essentially never worse than either baseline $\lambda = 1$ or $\lambda = 0$. When the noise is low ($\sigma_y^2 = 0.1$), greedy $\hat{\lambda}$ behaves similarly to $\lambda = 0$, placing almost all weight on the true labels. Conversely, when the noise is high ($\sigma_y^2 = 0.4$), greedy $\hat{\lambda}$ behaves like $\lambda = 1$, relying more heavily on pseudo-labels to reduce variance. For the intermediate noise level $\sigma_y^2 = 0.2$ in Figure 1, greedy $\hat{\lambda}$ significantly outperforms both baselines by hitting the optimal bias–variance balance.

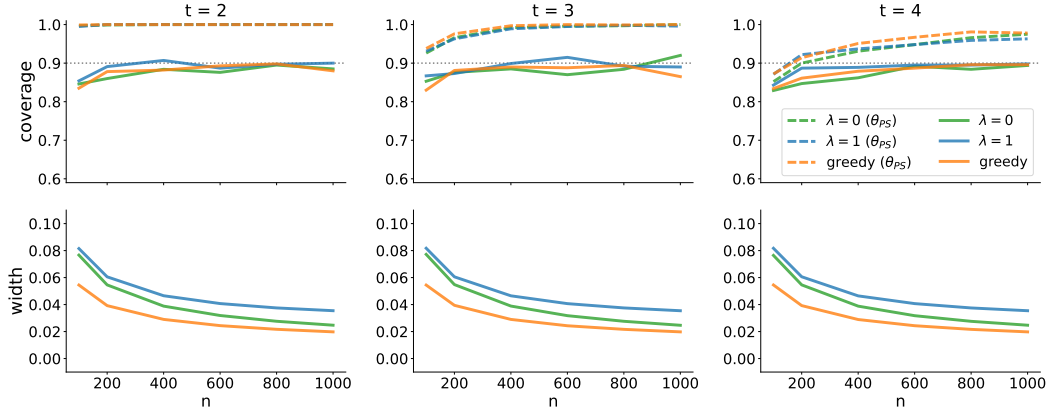
B.3 Case Study on Semi-synthetic Dataset

In Section 5, we originally consider experiments on a synthetic dataset because the performative prediction is an on-policy setting, which means we need to collect the corresponding data every time we update the parameter (policy). In all the previous literature on performative prediction, no such dataset is provided. Alternatively, previous work always uses a **semi-synthetic** dataset, which one will need to specify how the data distribution will react and shift according to the new policy.

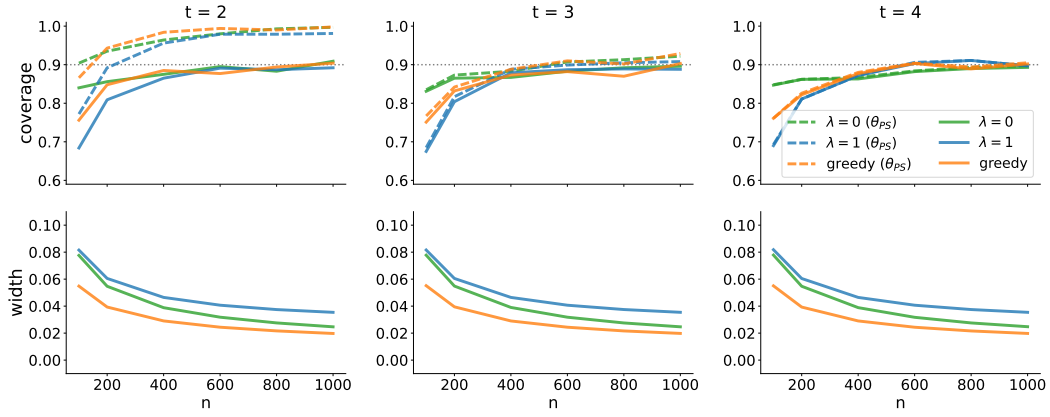
Following Perdomo et al. [25], we further conduct a case study in a semi-synthetic way on a realistic credit scoring task using a Kaggle dataset ⁴. The dataset contains features of individuals and a binary label indicating whether a loan should be granted or not. Consider the setting where a bank uses a logistic regression classifier θ trained on features of loan applicants to predict their creditworthiness, while the individual applicants respond to this classifier by manipulating their features to induce a positive classification. Following [2], we can formulate this task as performative prediction because applicants’ feature distribution $\mathcal{D}(\theta)$ is strategically adapted in response to θ . By applying repeated risk minimization, a performative stable point θ_{PS} can be achieved.

We treat the data points in the original dataset as the true distribution to compute θ_{PS} . We add Gaussian noise to the original data feature to generate an unlabeled set of the same size. Then, we sample varying n labeled points with $N = 18000$ unlabeled points and perform $t = 5$ repeated risk minimization steps to compute the estimated $\hat{\theta}_t$ and build the confidence region for it over 100 independent trials. From the experimental results, we find that a coverage of 0.9 is achieved with decreasing width as n increases. Notably, our optimized greedy $\hat{\lambda}_t$ (orange) achieves the highest coverage and narrowest confidence width compared with when λ is fixed to 0 or 1. The results support our proposed theory and strengthen the practical significance of our methods. We hope this case study can inspire future work for practical settings of PPI under performativity.

⁴<https://www.kaggle.com/c/GiveMeSomeCredit/data>

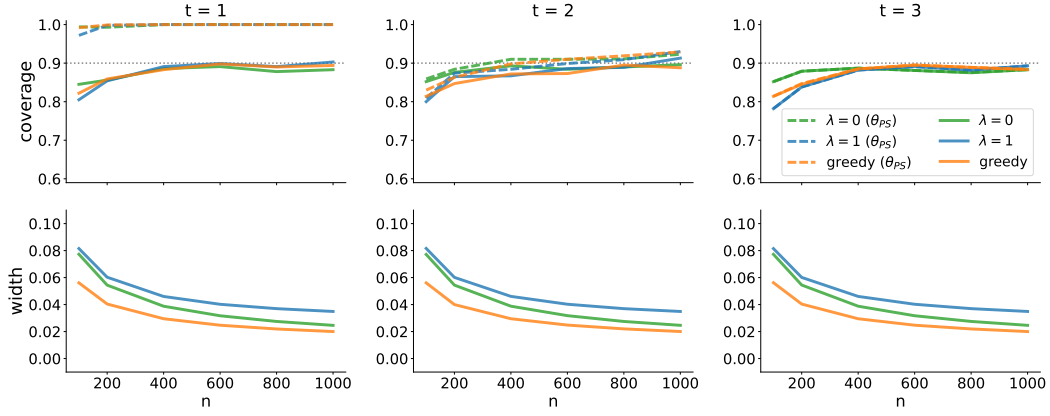


(a) Confidence-region coverage and width with $\gamma = 1$.

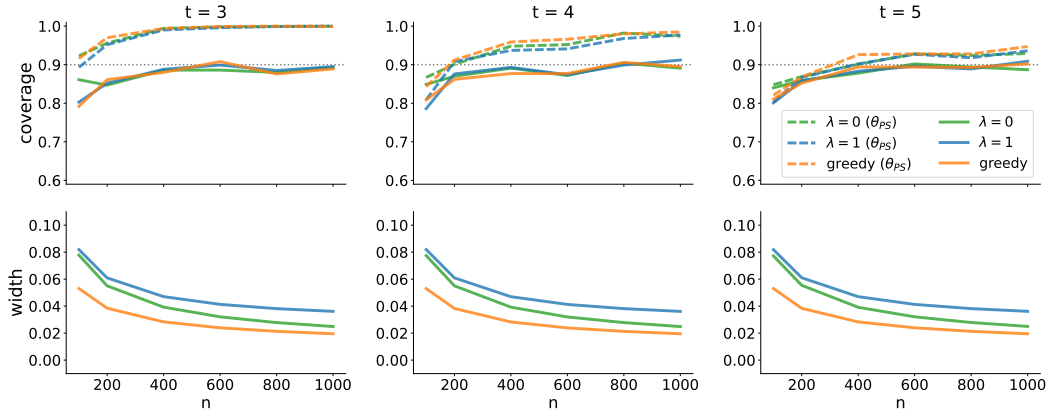


(b) Confidence-region coverage and width with $\gamma = 3$.

Figure A1: Confidence-region coverage (top row) and width (bottom row) with different choices of λ . The setup is the same as in Figure 1, only we change $\gamma = 1$ or $\gamma = 3$.

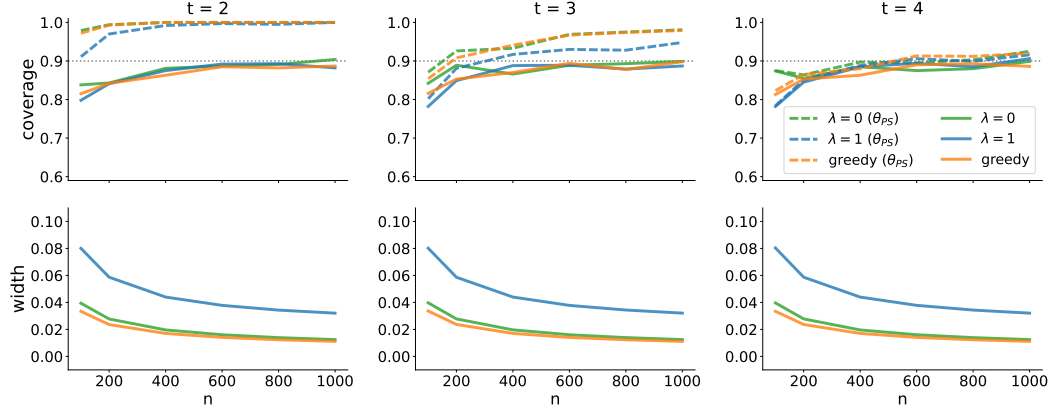


(a) Confidence-region coverage and width with $\varepsilon \approx 0.003$.

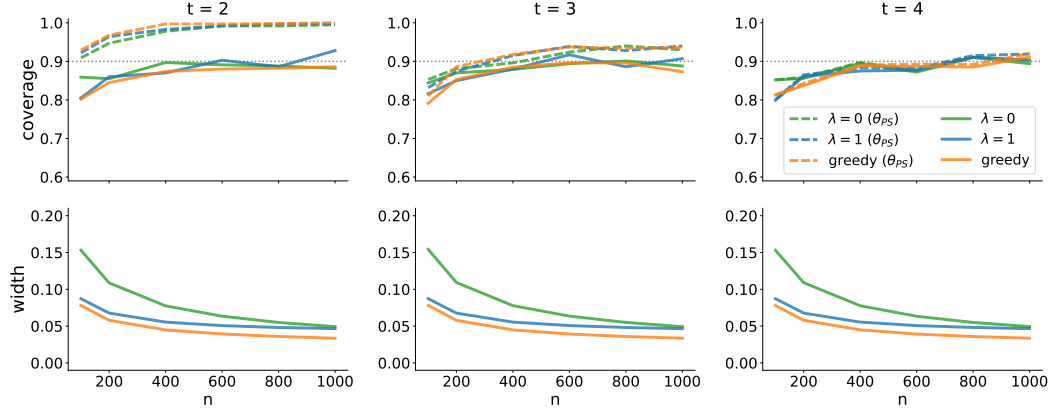


(b) Confidence-region coverage and width with $\varepsilon \approx 0.03$.

Figure A2: Confidence-region coverage (top row) and width (bottom row) with different choices of λ . The setup is the same as in Figure 1, only we change $\varepsilon \approx 0.003$ or $\varepsilon \approx 0.03$.



(a) Confidence-region coverage and width with $\sigma_y^2 = 0.1$.



(b) Confidence-region coverage and width with $\sigma_y^2 = 0.4$.

Figure A3: Confidence-region coverage (top row) and width (bottom row) with different choices of λ . The setup is the same as in Figure 1, only we change $\sigma_y^2 = 0.1$ or $\sigma_y^2 = 0.4$.

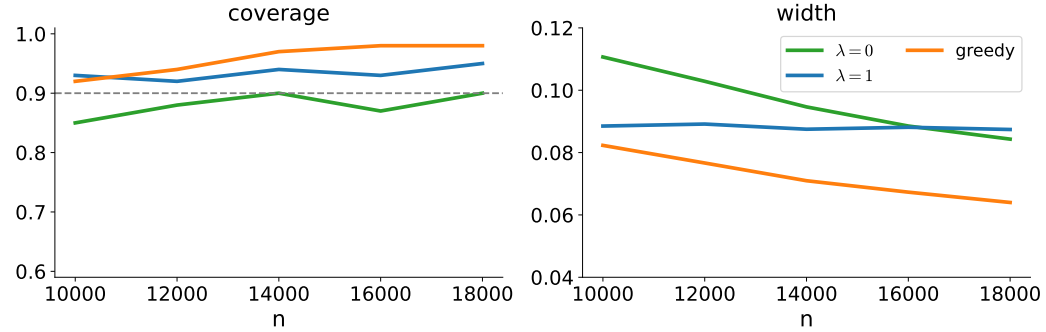


Figure A4: Confidence-region coverage and width for θ_t ($t = 5$) with different choices of λ on the semi-synthetic Kaggle credit scoring dataset.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [NA].
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading “NeurIPS Paper Checklist”,**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: The main claims in abstract and introduction are included in our paper. Specifically, our main theoretical results on CLT is included in Section 3. The results with prediction-powered inference is included in Section 4. We further demonstrate our numerical experiments in Section 5.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Currently, our framework uses bias-awared inference for θ_{PS} rather than a direct inference method. We hope that in the future, we could Obtain direct inference methods for θ_{PS} .

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We have listed our necessary assumptions in Assumption 3.1 and in Section A.1 in Appendix A. We also provide complete proofs of our claims in Appendix A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We explain our experimental settings and provide the complete set of hyperparameters in Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide a zip file containing our simulation study’s code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: We provide information about the training details in 5 and Appendix B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[No\]](#)

Justification: We do not provide error bars as the main results in Figure 1 capture the statistical significance of the experiments by counting the coverage rate of the confidence interval from repeated experimental trials. The randomness comes from the data generation process during performantive inference.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: All these information can be found in Appendix B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research conducted in this paper conforms, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: As we mentioned in our introduction, our paper is mainly for inference theory under performativity, which is quite useful in social science and provide a useful tool when evaluating policies made via performative prediction. There is no negative societal impacts we can think of.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not release any datasets or models since we are doing synthetic data experiments on very classical models like regression models.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: We do not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We provide code and instructions on how to run the code to reproduce our results.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.