Stochastic Optimization Schemes for Performative Prediction with Nonconvex Loss

Qiang Li, Hoi-To Wai

Dept of System Engineering and Engineering Management,
The Chinese University of Hong Kong

December 1, 2024
NeurIPS 2024, Vancouver, Canada

Overview

Performative Prediction

- **Motivation**: Learning in economic or societal environment is causative: the models aim to predict can be influenced by the models themselves.

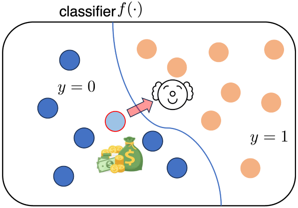
Examples:

- **Hiring process**: Job Description $\rightarrow$ applicants tailor their CV $\rightarrow$ Employer evaluates applicants.
  - Applicants who prepared strategically have an advantage, improving their chances of being hired.
- **Spam Email Detection**: Email server design filters to protect their users $\rightarrow$ Spammers circumvent filters to distribute malware and ADs.

Performative Prediction (Cont'd)

Loan application scenario:

- Bank's Approval Criteria $f(\cdot)$
- Denied Applicant's Response
- Strategic Adaptation
- Increased Chances.



Applicants' behavior:

- 1. know ...
- 2. want?
- 3. do!
- 4. outcome

- **Two Entities**: learner and agents' population.
- **Key Difference** between classical supervised learning: **intelligent agent's behavior**.

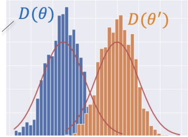
Mathematical Model

Perdomo et al. (2020) proposed to study the risk minimization problem with a decision-dependent data distribution:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} V(\boldsymbol{\theta}) := \mathbb{E}_{Z \sim \mathcal{D}(\boldsymbol{\theta})}[\ell(\boldsymbol{\theta}; Z)] \tag{1}$$

where $\ell(\boldsymbol{\theta}; Z)$ is continuously differentiable loss function *w.r.t.* $\boldsymbol{\theta}$ for given $z \in$ Z.

- Model entire population's responses.
- Avoids micro-level agent incentive modeling.
- ϵ-sensitive assumption: $d(\mathcal{D}(\boldsymbol{\theta}), \mathcal{D}(\boldsymbol{\theta}')) \leq \epsilon \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$.



Position of Perf. Pred.:

- Performative Prediction is a special example of distribution shift and causality, it lies in the intersection between machine learning and game theory.
- Another lines investigate distribution shift is strategic machine learning, see Rosenfeld (2024).

Research Gap

- Existing analysis are limited to the case when $\ell(\boldsymbol{\theta}; Z)$ are strongly convex *w.r.t.* $\boldsymbol{\theta}$ or impose structure on $\mathcal{D}(\boldsymbol{\theta})$.

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} V(\boldsymbol{\theta}) := \mathbb{E}_{Z \sim \mathcal{D}(\boldsymbol{\theta})}[\ell(\boldsymbol{\theta}; Z)] \rightarrow \text{scvx}$$

- Perdomo et al. (2020) introduced *performative stable* (PS) solution as the unique minimizer of (1) with fixed dist., i.e.,

$$\boldsymbol{\theta}_{PS} := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathbb{E}_{Z \in \mathcal{D}(\boldsymbol{\theta}_{PS})}[\ell(\boldsymbol{\theta}; Z)] \quad \rightarrow \text{fixed point sol.}$$

- **Algorithm**: SGD with greedy deployment recursion:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_{t+1} \nabla \ell(\boldsymbol{\theta}_t; Z_{t+1}), \text{ where } Z_{t+1} \sim \mathcal{D}(\boldsymbol{\theta}_t) \qquad (2)$$

- **Cons**: strong convexity assumption limits the class of classifier in machine learning tasks, such as neural network.

- In non-convex analysis, we need a new metric $\rightarrow$ SPS solution.

Our Contribution

- We firstly propose the concept of stationary performative stable (SPS) solution relaxing PS condition, which is necessary for handling non-convex losses using first-order methods.
- δ **stationary performative stable (SPS) solution**: Let $\delta \geq 0$, the vector $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ is said to be an δ stationary performative stable (δ-SPS) solution if:

$$\|\nabla_1 J(\boldsymbol{\theta}^\star; \boldsymbol{\theta}^\star)\|^2 = \left\|\mathbb{E}_{Z \sim \mathcal{D}(\boldsymbol{\theta}^\star)} \left[\nabla \ell(\boldsymbol{\theta}^\star; Z)\right]\right\|^2 \leq \delta.$$

- $\delta \geq 0$ measures the stationarity of a solution.
- If $\ell(\boldsymbol{\theta}; z)$ is strongly convex w.r.t. $\boldsymbol{\theta}$, then an SPS solution is also a PS solution.

Our Contribution (Cont'd)

- We show that SGD-GD finds a $\mathcal{O}(\epsilon)$-biased SPS solution.
  - Bias level is further improved to $\mathcal{O}(\epsilon^2)$ when the gradient is exact.
- **Techniques**: our analysis relies on constructing a time varying Laypunov function.
  - We study two alternative conditions on the distance metric: Wasserstein-1 distance and total variation (TV) distance.
- **Extension**: we extend the analysis to the lazy deployment scheme with SGD. As the epoch length grows, it can find bias-free SPS solution.

Definitions & Assumptions

$$J(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2) = \mathbb{E}_{Z\sim\mathcal{D}(\boldsymbol{\theta}_2)}\left[\ell(\boldsymbol{\theta}_1;Z)\right], \quad \nabla_1 J(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2) = \mathbb{E}_{Z\sim\mathcal{D}(\boldsymbol{\theta}_2)}\left[\nabla\ell(\boldsymbol{\theta}_1;Z)\right].$$

We observe that $V(\boldsymbol{\theta}) = J(\boldsymbol{\theta},\boldsymbol{\theta})$, $\nabla V(\boldsymbol{\theta}) \neq \nabla_1 J(\boldsymbol{\theta};\boldsymbol{\theta})$ in general.

- **A1**: $\|\nabla\ell(\boldsymbol{\theta};z)-\nabla\ell(\boldsymbol{\theta}';z)\| \leq L\,\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|, \forall\boldsymbol{\theta},\boldsymbol{\theta}' \in \mathbb{R}^d$, $\ell(\boldsymbol{\theta};z) \geq \ell^\star > -\infty$.
- **A2**: Assume that there exists constants $\sigma_0, \sigma_1 \geq 0$ such that
$$\mathbb{E}_{Z\sim\mathcal{D}(\boldsymbol{\theta}_2)}\left[\left\|\nabla\ell(\boldsymbol{\theta}_1;Z) - \nabla_1 J(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2)\right\|^2\right] \leq \sigma_0^2 + \sigma_1^2\left\|\nabla J(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2)\right\|^2.$$
- **A3**: ϵ sensitivity $d(\mathcal{D}(\boldsymbol{\theta}),\mathcal{D}(\boldsymbol{\theta}')) \leq \epsilon\,\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|$. (will be specified later.)

Main Results (I)

Theorem 1. Under A1-3. Let the step size satisfies $\sup_{t\geq 1}\gamma_t \leq 1/(L(1+\sigma_1^2))$. Then, for any $T \geq 1$, the iterates $\{\boldsymbol{\theta}_t\}_{t\geq 0}$ generated by SGD-GD satisfy

$$\sum_{t=0}^{T-1}\frac{\gamma_{t+1}}{4}\mathbb{E}\,\|\nabla_1 J(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t)\|^2 \leq \Delta_0 + \tilde{L}\epsilon\left(\sigma_0 + (1+\sigma_1^2)\tilde{L}\epsilon\right)\sum_{t=0}^{T-1}\gamma_{t+1} + \frac{L\sigma_0^2}{2}\sum_{t=0}^{T-1}\gamma_{t+1}^2,$$

where $\Delta_0 := J(\boldsymbol{\theta}_0;\boldsymbol{\theta}_0) - \ell_\star$ is an upper bound to the initial optimality gap of performative risk.

- **Corollary 1**. Under A1-3. Let $T \geq 1$ be the maximum number of iterations and set $\gamma_t = 1/\sqrt{T}$. For any sufficient large T, the iterates by SGD-GD satisfy

$$\mathbb{E}\left[\|\nabla_1 J(\boldsymbol{\theta}_\mathsf{T};\boldsymbol{\theta}_\mathsf{T})\|^2\right] \leq 4\left(\Delta_0 + \frac{L}{2}\sigma_0^2\right)\cdot\frac{1}{\sqrt{T}} + \underbrace{4\tilde{L}\epsilon\left(\sigma_0 + (1+\sigma_1^2)\tilde{L}\epsilon\right)}_{\mathcal{O}(\epsilon\sigma_0+\epsilon^2)-\textbf{bias}}. \quad (3)$$

where T is a r.v. chosen uniformly and independently from $\{0, 1, \cdots, T-1\}$.

Discussion of Theorem 1

- **Corollary 1**. Set $\gamma_t = 1/\sqrt{T}$. For sufficient large T, it holds that

$$\mathbb{E}[\|\nabla_1 J(\boldsymbol{\theta}_{\mathsf{T}}; \boldsymbol{\theta}_{\mathsf{T}})\|^2] \lesssim \frac{1}{\sqrt{T}} + \widetilde{L}\left(\epsilon\sigma_0 + \epsilon^2\right).$$

- SGD-GD finds a $\mathcal{O}(\epsilon)$-biased SPS solution.
- Bias level is further improved to $\mathcal{O}(\epsilon^2)$-biased when the gradient is exact.
- The asymptotic performance of SGD-GD is sensitive to the stochastic gradient's noise variance.

Key Lemma I: Descent Lemma

Lemma 1. Under A1, 2. Suppose that the step size satisfies

$$\sup_{t \geq 1} \gamma_t \leq 1/(L(1+\sigma_1^2)),$$

then for any $t \geq 0$, the iterates generated by SGD-GD satisfies

$$\frac{\gamma_{t+1}}{2} \|\nabla_1 J(\boldsymbol{\theta}_t, \boldsymbol{\theta}_t)\|^2 \leq \underbrace{J(\boldsymbol{\theta}_t, \boldsymbol{\theta}_t) - \mathbb{E}_t[J(\boldsymbol{\theta}_{t+1}; \boldsymbol{\theta}_t)]}_{:= A_1} + \frac{L\sigma_0^2}{2} \gamma_{t+1}^2. \quad (4)$$

- For sufficiently small γ_t and when $\boldsymbol{\theta}_t$ is not SPS, (4) implies the descent relation

$$\mathbb{E}_t[J(\boldsymbol{\theta}_{t+1}; \boldsymbol{\theta}_t)] \leq J(\boldsymbol{\theta}_t; \boldsymbol{\theta}_t)$$

- Motivated by above relation, we consider $J(\boldsymbol{\theta}_t; \boldsymbol{\theta}_t)$ as the time-varying Laypunov function.

$$\mathbb{E}[A_1] = \mathbb{E}[J(\boldsymbol{\theta}_t; \boldsymbol{\theta}_t) - J(\boldsymbol{\theta}_{t+1}; \boldsymbol{\theta}_{t+1})] + \underbrace{\mathbb{E}[J(\boldsymbol{\theta}_{t+1}; \boldsymbol{\theta}_{t+1}) - J(\boldsymbol{\theta}_{t+1}; \boldsymbol{\theta}_t)]}_{\text{residual}}$$

- **Question**: How to bound residual term?

Key Lemma II: Bound Distribution Shift

Recall the distribution smooth assumption: $d(\mathcal{D}(\boldsymbol{\theta}), \mathcal{D}(\boldsymbol{\theta}')) \leq \epsilon \left\| \boldsymbol{\theta} - \boldsymbol{\theta}' \right\|$.

- **W1**: ϵ **sensitivity** $\mathcal{W}_1(\mathcal{D}(\boldsymbol{\theta}), \mathcal{D}(\boldsymbol{\theta}')) \leq \epsilon \left\| \boldsymbol{\theta} - \boldsymbol{\theta}' \right\|$.
- **W2**: L_0 **smoothness w.r.t. sample** $|\ell(\boldsymbol{\theta}; z) - \ell(\boldsymbol{\theta}; z'))| \leq L_0 \left\| z - z' \right\|$.

W1 is standard, but **W2** can be difficult to verify.

- **C1**: ϵ **sensitivity** $\delta_{\mathsf{TV}}(\mathcal{D}(\boldsymbol{\theta}_1), \mathcal{D}(\boldsymbol{\theta}_2)) \leq \epsilon \left\| \boldsymbol{\theta} - \boldsymbol{\theta}' \right\|$.
- **C2**: **bounded loss** $\sup_{\boldsymbol{\theta} \in \mathbb{R}^d, z \in \mathsf{Z}} |\ell(\boldsymbol{\theta}; z)| \leq \ell_{\max}$.

C1 is slightly strengthened from **W1**. But **C2** covers more loss functions.

- **Lemma 2**. For any $\boldsymbol{\theta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$, it holds

$$|J(\boldsymbol{\theta}; \boldsymbol{\theta}_1) - J(\boldsymbol{\theta}; \boldsymbol{\theta}_2)| \leq \widetilde{L}\epsilon \left\| \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\| \tag{5}$$

  Under **W1 & 2**, $\widetilde{L} = L_0$, Under **C1 & 2**, $\widetilde{L} = 2\ell_{\max}$.

Combined Lemmas 1 & 2, we can obtained the Theorem 1.

Main Result (II) – Extension: Lazy Deployment with SGD

As inspired by Mendler-Dünner et al. (2020), lazy deployment scheme is described as following,

$$\begin{aligned}&\boldsymbol{\theta}_{t,k+1} = \boldsymbol{\theta}_{t,k} - \gamma \nabla \ell(\boldsymbol{\theta}_{t,k}; Z_{t,k+1}), \text{ where } Z_{t,k+1} \sim \mathcal{D}(\boldsymbol{\theta}_t), \\ &\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_{t+1,0} = \boldsymbol{\theta}_{t,K}, \quad k = 0, ..., K-1.\end{aligned} \tag{6}$$

Theorem 2. Under A1-3, and suppose that $\sup_{\boldsymbol{\theta} \in \mathbb{R}^d, z \in \mathsf{Z}} \|\nabla \ell(\boldsymbol{\theta}; z)\| \leq G$. Set $\gamma = 1/(K\sqrt{T})$. For sufficient large T, it holds that

$$\mathbb{E}\left[\|\nabla_1 J(\boldsymbol{\theta}_\mathsf{T}; \boldsymbol{\theta}_\mathsf{T})\|^2\right] \lesssim \frac{\Delta_0}{\sqrt{T}} + \frac{L\sigma_0^2}{K\sqrt{T}} + \frac{LG^2}{T} + \frac{\tilde{L}\epsilon}{K}\left(\sqrt{K}\sigma_0 + \sqrt{(K+\sigma_1^2)}\tilde{L}\epsilon\right).$$

After simplification, we have

$$\mathbb{E}\left[\|\nabla_1 J(\boldsymbol{\theta}_\mathsf{T}; \boldsymbol{\theta}_\mathsf{T})\|^2\right] \lesssim \mathcal{O}\left(\frac{1}{\sqrt{T}} + \frac{\widetilde{L}\epsilon}{\sqrt{K}}\right) \tag{7}$$

- The lazy deployment scheme (6) finds a *bias-free SPS solution* when $T \to \infty, K \to \infty$.

Numerical Experiments - Synthetic Data

Synthetic Data with Linear Model. We consider a binary classification problem with linear model,

$$\ell(\boldsymbol{\theta}; z) := (1 + \exp(c \cdot y\langle x \,|\, \boldsymbol{\theta}\rangle))^{-1} + (\beta/2) \left\|\boldsymbol{\theta}\right\|^2 ,$$

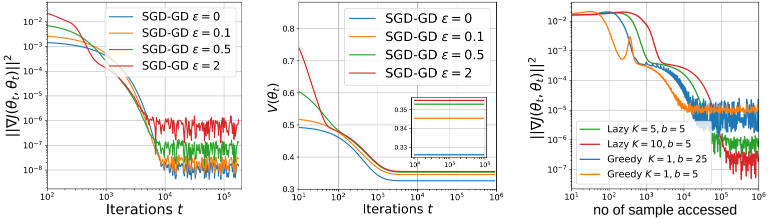
for small regularization $\beta > 0$, $\ell(\cdot; z)$ is smooth but non-convex.

Generating data distribution: $\mathcal{D}^o \equiv \{(x_i, y_i)\}_{i=1}^m$ with d-dimention feature $x_i \sim \mathcal{U}[-1, 1]^d$ and label $y_i = \mathrm{sgn}(\langle x_i \,|\, \boldsymbol{\theta}^o\rangle) \in \{\pm 1\}$, such that $\boldsymbol{\theta}^o \sim \mathcal{N}(0, \boldsymbol{I})$.

Distribution Shift: For any $\boldsymbol{\theta} \in \mathbb{R}^d$, $\mathcal{D}(\boldsymbol{\theta})$ is a uniform distribution on m *shifted samples* $\{(x_i - \epsilon_L \boldsymbol{\theta}, y_i)\}_{i=1}^m$, where $\epsilon_L > 0$ controls shift magnitude.

Parameter Set. $m = 800, d = 10, c = 0.1, \beta = 10^{-3}, \epsilon \in \{0, 0.1, 0.5, 2\}$. For SGD-GD, batch size: $b = 1$, stepsize: $\gamma_t = \gamma = 1/\sqrt{T}$ with $T = 10^6$.

Simulation Result - Synthetic Data



- From left figure, After a rapid transient stage, the SPS stationarity $\|\nabla J(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t)\|^2$ saturates and stay around a constant level, indicating that the SGD-GD converges to a biased-SPS solution. $\epsilon_L \uparrow$ leads to an increased bias. $\rightarrow$ **Theorem 1** ✓
- In middle figure, we evaluate the performance of the trained classifier $\boldsymbol{\theta}_t$ in terms of the performative risk value $V(\boldsymbol{\theta}_t)$.
- In right figure, we compare the lazy deployment with $K \in \{5, 10\}$ and stepsize $\gamma = 1/(K\sqrt{T})$. $K \uparrow$ leads to lower bias. $\rightarrow$ **Theorem 2** ✓

Numerical Experiments - Real Data

Spam Email Classification with Neural Network(NN) Model

- **Dataset**: Hopkins et al. (1999) with $m = 4601$ samples, $d = 48$ features. Training/test set: 8:2. Label $y \in \{0, 1\}$ (0 is for not spam, 1 for spam). Denote unshifted data as $\mathcal{D}^o = \{(\bar{x}_i, \bar{y}_i)\}_{i=1}^m$.
- **Problem formulation**: Consider the regularized binary cross entropy loss:

$$\begin{aligned} \ell(\boldsymbol{\theta}; z) &\equiv \tilde{\ell}(f_{\boldsymbol{\theta}}(x); y) \\ &= -y \log(f_{\boldsymbol{\theta}}(x)) - (1-y) \log(1 - f_{\boldsymbol{\theta}}(x)) + (\beta/2) \|\boldsymbol{\theta}\|^2, \end{aligned} \tag{8}$$

  where $f_{\boldsymbol{\theta}}(x)$ is the NN classifier.
- **Distribution Shift**: drawn new sample via maximizing the utility function:

$$x = \arg\max_{x'} U(x'; \bar{x}, \boldsymbol{\theta}) := -f_{\boldsymbol{\theta}}(x') - \frac{1}{2\epsilon_{\mathsf{NN}}} \|x' - \bar{x}\|^2, \tag{9}$$

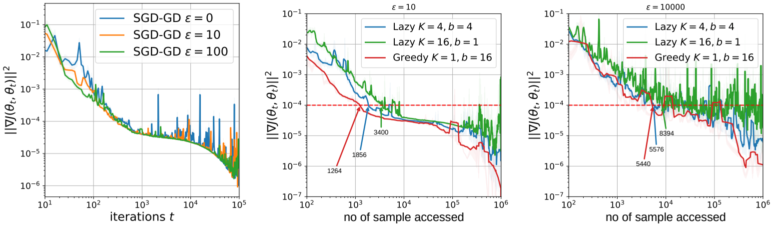
  to get $z \equiv (x, \bar{y}) \sim \mathcal{D}(\boldsymbol{\theta})$. In practice, we take approx. $x \approx \bar{x} - \epsilon_{\mathsf{NN}} \nabla_x f_{\boldsymbol{\theta}}(\bar{x})$.
- **NN Classifier**: three fully-connected layers with $\tanh$ activation and a sigmoid output layer,

$$f_{\boldsymbol{\theta}}(x) = Sigmoid\big(\boldsymbol{\theta}_{(1)}^\top \cdot \tanh(\boldsymbol{\theta}_{(2)}^\top \cdot \tanh(\boldsymbol{\theta}_{(3)}^\top x))\big),$$

  where $\boldsymbol{\theta}_{(i)} := [w_{(i)}; b_{(i)}] \in \mathbb{R}^{3421}$ concatenates the weight and bias.

Simulation Result - Real Data

- **Settings**: $\epsilon_{\mathsf{NN}} \in \{0, 10, 100\}$, batch size $b = 8$. For SGD-GD: $\gamma_t = \gamma = 200/\sqrt{T}$, Lazy deployment, $\gamma = 200/(K\sqrt{T})$ with $T = 10^5$.



Observation:

- From left fig, SGD-GD converges to a near SPS solution.
- From middle & right fig, lazy deployment performs relatively better than SGD-GD as $\epsilon_{\mathsf{NN}} \uparrow$.

  When $\epsilon : 10 \mapsto 10^5$, no. sample for three algo: $\times 4$, $\times 3$, $\times 2.4$.

- Recall from (7), $\mathbb{E}[\|\nabla_1 J(\boldsymbol{\theta}_{\mathsf{T}}; \boldsymbol{\theta}_{\mathsf{T}})\|^2] = \mathcal{O}\left(\frac{\epsilon}{\sqrt{K}}\right)$ and $\epsilon \propto \epsilon_{\mathsf{NN}}$.

Conclusions

- We provides the first study on the performative prediction problem with **smooth but possibly non-convex** loss.
- A stationary performative stable (SPS) condition which is the counterpart of performative stable condition used with strongly convex loss, is developed to analyze nonconvex case.
- We provide the convergence of greedy deployment and lazy deployment schemes with SGD under nonconvex case.
- Numerical experiments validate our analysis.

- **Limitation/ongoing work**: Nonconvex analysis based on non-iid data?

Questions & Comments?

References I

Hopkins, M., Reeber, E., Forman, G., and Suermondt, J. (1999). Spambase. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C53G6X.

Mendler-Dünner, C., Perdomo, J., Zrnic, T., and Hardt, M. (2020). Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33:4929–4939.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. (2020). Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR.

Rosenfeld, N. (2024). Strategic ml: How to learn with data that 'behaves'. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 1128–1131, New York, NY, USA. Association for Computing Machinery.