Stochastic Optimization Schemes for Performative Prediction with Nonconvex Loss

Informs International Meeting 2025

Qiang Li

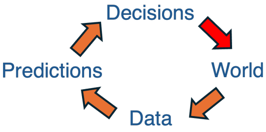
Supervisor: Prof. Hoi-To Wai

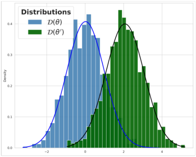
Dept of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong

July 21, 2025

Motivation: Data Distribution May Shift

- **Performative Prediction (PP)**: stochastic optimization problem whose data distribution depends on the decision variable.





- Learning in economic or societal environment is causative: the models aim to predict can be influenced by the models themselves.
  - Example: self-fulfilling or self-negating predictions.

- **Example (I)**: Spam Email Detection
  - An email server designs a filter to block spam.
  - Spammers *adapt to bypass* the filter and continue distributing spam or malware.
- **Example (II)**: Traffic Congestion
  - Google Maps suggests the fastest route based on current traffic conditions.
  - Many users follow the suggestion, the recommended route becomes congested.
- **Related topic**: Stackberg games (Brückner and Scheffer, 2011).

From Practice to Mathematical Model

- **Performative Prediction**: Data $Z = (x, y) \sim \mathcal{D}(\boldsymbol{\theta})$
- **Formulation**: minimize the performative risk

$$\min_{\boldsymbol{\theta}} V(\boldsymbol{\theta}) := \mathbb{E}_{Z \sim \mathcal{D}(\boldsymbol{\theta})}[\ell(\boldsymbol{\theta}; Z)]$$

- **Example** of $\mathcal{D}(\boldsymbol{\theta})$: base distribution $\mathcal{D}^o \equiv \{(x_i, y_i)\}_{i=1}^m$, (x_i, y_i) is feature label pair, $\mathcal{D}(\boldsymbol{\theta}) = \{(x_i - \epsilon\boldsymbol{\theta}, y_i)\}_{i=1}^m$, where ϵ is shift magnitude.

- Perdomo et al. (2020) uses $\mathcal{D}(\boldsymbol{\theta})$ to capture the distribution shift (population's response of Z) due to the learner's state $\boldsymbol{\theta}$.
- *How should the learner deal with performativity?*
  - Agnostic Setting: SGD with greedy deployment on $\ell(\boldsymbol{\theta}; z)$ with $z \sim \mathcal{D}(\boldsymbol{\theta})$, e.g., Perdomo et al. (2020), Mendler-Dünner et al. (2020).
  - Requires no extra knowledge on $V(\boldsymbol{\theta})$ and population ...

  - *Proactive Setting*: Estimate true gradient of $\nabla V(\boldsymbol{\theta})$, e.g., Izzo et al. (2021), Miller et al. (2021).
  - Needs extra knowledge on $V(\boldsymbol{\theta})$ and population utility function.

SGD with Greedy Deployment (Mendler-Dünner et al., 2020)

- Two different solutions to performative prediction:

$$\theta_{PO} \in \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{Z\sim\mathcal{D}(\theta)}[\ell(\theta; Z)], \quad \theta_{PS} \in \arg\min_{\theta' \in \mathbb{R}^d} \mathbb{E}_{Z\sim\mathcal{D}(\theta_{PS})}[\ell(\theta'; Z)].$$

- In **agnostic setting**, our aim is to get θ_{PS}, e.g., by fixed point iteration. How can we find it?

Greedy deployment scheme (Mendler-Dünner et al., 2020):

$$\underline{\textit{Population}}: \quad Z_{t+1} \sim \mathcal{D}(\theta_t),$$
$$\underline{\textit{Learner (Agent)}}: \quad \theta_{t+1} = \theta_t - \gamma_{t+1}\nabla\ell(\theta_t; Z_{t+1}).$$

- Illustration of SGD w/ GD at iteration t,

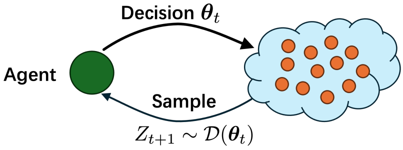


Figure 1: SGD with greedy deployment

SGD with Greedy Deployment (Cont'd)

W1: The distribution $\mathcal{D}(\boldsymbol{\theta})$ satisfies ϵ-sensitivity if for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$,

$$\mathcal{W}_1(\mathcal{D}(\boldsymbol{\theta}), \mathcal{D}(\boldsymbol{\theta}')) \leq \epsilon \left\| \boldsymbol{\theta} - \boldsymbol{\theta}' \right\|.$$

where $\mathcal{W}_1$ denotes Wasserstein-1 distance.

- **Fact I**: if $\ell(\cdot; Z)$ is strongly convex + $\mathcal{D}(\boldsymbol{\theta})$ is 'insensitive' to $\boldsymbol{\theta}$, then

$$\mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{PS}\|^2] = \mathcal{O}(1/t).$$

- **Fact II**: (Perdomo et al., 2020) Suppose that $\ell(\boldsymbol{\theta}; z)$ is L-smooth, μ-strongly convex and distribution $\mathcal{D}(\cdot)$ is ϵ-sensitive,

$$\|\boldsymbol{\theta}_{PS} - \boldsymbol{\theta}_{PO}\|_2 \leq \frac{2L\epsilon}{\mu}$$

Research Q: If $\ell(\boldsymbol{\theta}; Z)$ is smooth but possibly **non-convex**, will SGD/GD converge to fixed point solution $\boldsymbol{\theta}_{PS}$?

Overview of This Talk

Performative Prediction with Non-convex Loss

Perf. Pred. with Non-convex Loss

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

- Well-definedness: The loss function is lower bounded.
- PS solution may not be unique, so we need a relaxed condition.

Def: The solution $\boldsymbol{\theta}_{SPS}$ is called a δ-Stationary PS solution if it satisfies

$$\left\| \mathbb{E}_{Z \sim \mathcal{D}(\boldsymbol{\theta}_{SPS})}[\nabla \ell(\boldsymbol{\theta}_{SPS}; Z)] \right\| \leq \delta.$$

If $\ell(\cdot)$ is strongly convex, (0-)SPS=PS.

- The stochastic gradient $\nabla \ell(\boldsymbol{\theta}_t; Z_{t+1})$ is not a gradient nor unbiased, since

$$\nabla V(\boldsymbol{\theta}) = \nabla \int_{\mathsf{Z}} \ell(\boldsymbol{\theta}; z) p_{\mathcal{D}(\boldsymbol{\theta})} \mathrm{d}z$$
$$= \mathbb{E}_{z \sim \mathcal{D}(\boldsymbol{\theta})}[\nabla \ell(\boldsymbol{\theta}; z)] + \mathbb{E}_{z \sim \mathcal{D}(\boldsymbol{\theta})}[\ell(\boldsymbol{\theta}; z) \nabla_{\boldsymbol{\theta}} \log(p_{\mathcal{D}(\boldsymbol{\theta})}(z))]$$

- Denote $f(\boldsymbol{\theta}_1; \boldsymbol{\theta}_2) := \mathbb{E}_{Z \sim \mathcal{D}(\boldsymbol{\theta}_2)}[\ell(\boldsymbol{\theta}_1; Z)]$, $\nabla f(\boldsymbol{\theta}_1; \boldsymbol{\theta}_2) := \mathbb{E}_{Z \sim \mathcal{D}(\boldsymbol{\theta}_2)}[\nabla \ell(\boldsymbol{\theta}_1; Z)]$

Contributions: Two Alternative Assumption Sets

- **W1**: **(Wasserstein sensitivity)** $\forall \boldsymbol{\theta}, \boldsymbol{\theta}'$, $\mathcal{W}_1(\mathcal{D}(\boldsymbol{\theta}), \mathcal{D}(\boldsymbol{\theta}')) \leq \epsilon \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$.
- **W2**: **(Lipschitz loss)** $|\ell(\boldsymbol{\theta}; z) - \ell(\boldsymbol{\theta}; z'))| \leq L_0 \|z - z'\|$.

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

- Note that **C1** is stronger than **W1**, but **C2** is weaker than **W2**.
- **Fact**: As shown in (Gibbs and Su, 2002, Sec. 2),
$$\mathcal{W}_1(\mathcal{D}(\boldsymbol{\theta}), \mathcal{D}(\boldsymbol{\theta}')) \leq \mathrm{diam}(\mathsf{Z}) \cdot d_{TV}(\mathcal{D}(\boldsymbol{\theta}), \mathcal{D}(\boldsymbol{\theta}'))$$
where $\mathrm{diam}(\mathsf{Z}) := \sup_{z, z' \in \mathsf{Z}} \|z - z'\|$ denotes the diam of the sample space.
- Sigmoid loss satisfies **C2** with $\ell_{max} = 1$ but not **W2** expect $\|z\|$ is bounded.
- **Remark**: **W1&2** and **C1&2** are used to quantify the distribution shift effect on Lyapunov function in convergence analysis.

Main Theorem (I)

A1: The gradient map $\nabla \ell(\cdot;\cdot)$ is L-Lipschitz,

$$\|\nabla \ell(\boldsymbol{\theta}_1; z_1) - \nabla \ell(\boldsymbol{\theta}_2; z_2)\| \leq L\left(\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + \|z_1 - z_2\|\right)$$

A2: **(Variance)** For all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, there exists $\sigma_0, \sigma_1 \geq 0$ such that

$$\mathbb{E}_{Z \sim \mathcal{D}(\boldsymbol{\theta}_2)} \|\nabla \ell(\boldsymbol{\theta}_1; Z) - \nabla f(\boldsymbol{\theta}_1; \boldsymbol{\theta}_2)\|^2 \leq \sigma_0^2 + \sigma_1^2 \|\nabla f(\boldsymbol{\theta}_1; \boldsymbol{\theta}_2)\|^2$$

Theorem 1: Let **A1,2**. Suppose that the stepsize satisfy $\sup_{t \geq 1} \gamma_t \leq \frac{1}{L(1+\sigma_1^2)}$. Moreover, let

$$\tilde{L} = L_0 \text{ if } \textbf{W1, 2} \text{ hold, or } \tilde{L} = 2\ell_{max} \text{ if } \textbf{C1,2} \text{ hold.}$$

Then, for any $T \geq 1$, it holds that

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

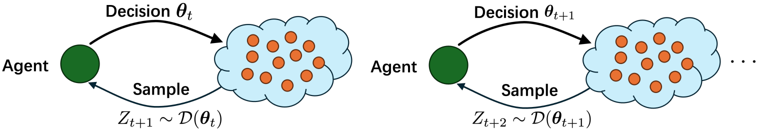
- If $\gamma_t = 1/\sqrt{T}$, then the iterates by SGD-GD satisfy

$$\mathbb{E}\left[\|\nabla f(\boldsymbol{\theta}_{\mathsf{T}}; \boldsymbol{\theta}_{\mathsf{T}})\|^2\right] \leq \mathcal{O}(1/\sqrt{T}) + \underbrace{4\tilde{L}\epsilon(\sigma_0 + (1+\sigma_1^2)\tilde{L}\epsilon)}_{=:\text{bias}}$$

- **Biased-SPS Solution**: $\mathcal{O}(\epsilon)$ for noisy SGD, $\mathcal{O}(\epsilon^2)$ for noiseless SGD.

Lazy Deployment

- **Greedy Deployment** $Z_t \sim \mathcal{D}(\boldsymbol{\theta}_t)$, requires deploying the latest model every time when drawing new samples from $\mathcal{D}(\cdot)$.

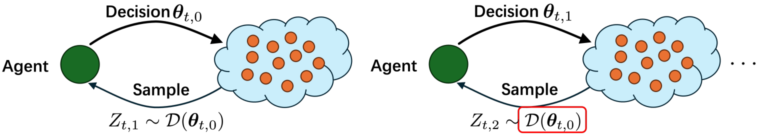


- The agent and population progress at the same pace.

Frequent deployment can be costly.

- **Lazy Deployment**: $K \geq 1$ denotes the epoch length,

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



- *The agent (learner) progresses faster than population $\rightarrow$ more accurate sol.*

Main Theorem (II) – Extension to Lazy Deployment

Theorem 2. Under **A1,2**, **W1,2** or **C1,2**, 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

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

where T is the random variable drawn from Unif$(\{1, 2, \cdots, T\})$.

After simplification, we have

$$\mathbb{E}\left[\left\|\nabla f(\boldsymbol{\theta}_{\mathsf{T}}; \boldsymbol{\theta}_{\mathsf{T}})\right\|^2\right] \lesssim \mathcal{O}\left(\frac{1}{\sqrt{T}} + (\tilde{L}\epsilon)^2 \frac{K + \sigma_1^2}{K}\right) \tag{1}$$

- Lazy deployment finds $\mathcal{O}(\epsilon^2)$-SPS solution, when $T, K \to \infty$, while SGD-GD finds $\mathcal{O}(\epsilon)$ solution.

Simulations - Binary Classification

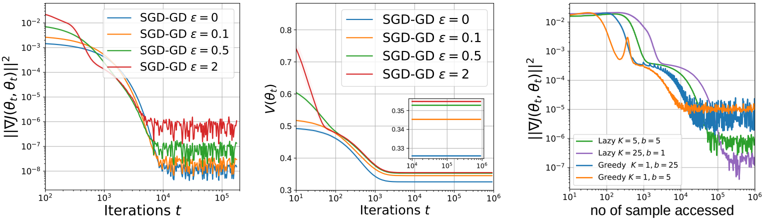
Synthetic Data with Linear Model.

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

for small regularization $\epsilon > 0$, $\ell(\cdot; z)$ 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 $y_i = \text{sgn}(\langle x_i \,|\, \boldsymbol{\theta}^o\rangle) \in \{\pm 1\}$, such that $\boldsymbol{\theta}^o \sim \mathcal{N}(0, \boldsymbol{I})$.

Dist. Shift: $\mathcal{D}(\boldsymbol{\theta}) = \text{Unif}\{(x_i - \epsilon_L \boldsymbol{\theta}, y_i)\}_{i=1}^m$, $\epsilon_L > 0$ controls shift magnitude.



- **Left & Middle Fig.**: SGD-GD shows a fast transient phase, then saturates near a constant; $\epsilon \propto$ bias $\rightarrow$ **Theorem 1** ✓
- **Right Fig.**: SGD-Lazy deployment with $K \in \{5, 10\}$ and stepsize $\gamma = 1/(K\sqrt{T})$. $K \uparrow$ leads to lower bias. $\rightarrow$ **Theorem 2** ✓

Conclusions

Performative Prediction

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

If $\ell(\boldsymbol{\theta}; Z)$ is smooth but possibly non-convex:

- **(A)** SGD with greedy deployment finds an $\mathcal{O}(\epsilon)$-biased SPS solution.
- **(B)** The bias can be reduced to $\mathcal{O}(\epsilon^2)$ with exact gradients.
- **(C)** SGD with lazy deployment yields a more accurate SPS solution as the episode length $\to \infty$.
- **Key idea**: use a time-varying Lyapunov function to analyze non-gradient dynamics.

Thank you for your time and attention!

Scan the qr code for the full paper $\to$

References I

Brückner, M. and Scheffer, T. (2011). Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 547–555.

Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435.

Izzo, Z., Ying, L., and Zou, J. (2021). How to learn when data reacts to your model: Performative gradient descent. In *ICML*.

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.

Miller, J. P., Perdomo, J. C., and Zrnic, T. (2021). Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, pages 7710–7720. PMLR.

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