# State-dependent Performative Prediction with Stochastic Approximation

Qiang LI, Hoi-To Wai Dept. of SEEM, The Chinese University of Hong Kong

## Prediction / ML

- Supervised learning: static + i.i.d. data.
- Decision can cause distribution shift.
- Performative Prediction: data distribution depends on decision variables.

**Goal:** minimize performative risk

$$\min_{\theta} \mathcal{L}(\theta) := \mathbb{E}_{z=(x,y)\sim\mathcal{D}(\theta)}[\ell(\theta; z)]$$

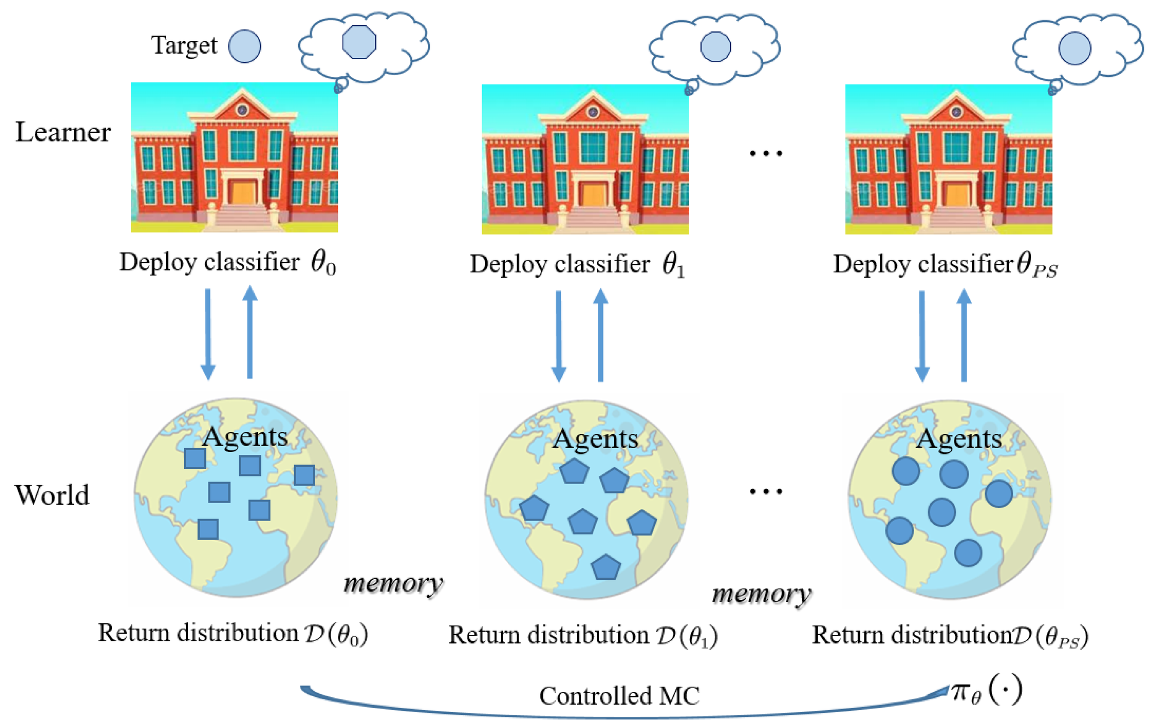
**Sol.**: $\theta_{PO} \in \arg\min_{\theta\in\mathbb{R}^d} \mathbb{E}_{z\sim\mathcal{D}(\theta)}\ell(\theta; z)$,
$\theta_{PS} = \arg\min_{\theta'\in\mathbb{R}^d} \mathbb{E}_{z\sim\mathcal{D}(\theta_{PS})}[\ell(\theta'; z)]$.

## What should the learner do?

- Proactive: estimate true gradient $\nabla\mathcal{L}(\theta)$.
- Agnostic: SGD/GD on $\ell(z;\theta)$ with $z \sim \mathcal{D}(\theta)$ (no extra knowledge on agents).
- Greedy Deploy [Mendler-Dünner, 2020]:
  - *learner*: $\theta_{k+1} = \theta_k - \gamma_{k+1}\nabla\ell(\theta_k; z_{k+1})$,
  - *agents*: $z_{k+1} \sim \mathcal{D}(\theta_k)$
  - Immediate vs slow adaptation.

## Highlights

- **State dependent** performative prediction framework.
- Agnostic scheme has $\theta_k \to \theta_{PS}$ at $\mathcal{O}(\frac{1}{k})$.




## State-dependent Performative Prediction with SA

- **Idea**: models agents' adaptation via a controlled Markov Chain described by kernel $\mathrm{P}_\theta : \mathsf{Z} \times \mathcal{Z} \to \mathbb{R}_+$ and stationary dist. $\mathcal{D}(\theta)$.

**SA Algorithm with Adapted Agent Response**

Agents draw: $z_{k+1} \sim \mathrm{P}_{\theta_k}(z_k, \cdot)$ (← allows slow adaptation)
Learner updates: $\theta_{k+1} = \theta_k - \gamma_{k+1}\nabla\ell(\theta_k; z_{k+1})$ and deploys $\theta_{k+1}$.

- **Example:** AR model, agents running SGD to adapt to $z \sim \mathcal{D}(\theta)$:

$$z_{k+1} = z_k + \alpha\nabla_z U(z_k; \theta_k, \zeta_{k+1}), \quad \leftarrow U = \text{utility fct.}$$

- **Challenge:** Analyze MSE $\mathbb{E}[\|\theta_k - \theta_{PS}\|^2]$ **w.o.** boundedness assumption.

## Main Results

- Define: $f(\theta_1; \theta_2) = \mathbb{E}_{z\sim\mathcal{D}(\theta_2)}[\ell(\theta_1; z)]$, $\nabla f(\theta_1; \theta_2) = \mathbb{E}_{z\sim\mathcal{D}(\theta_2)}[\nabla\ell(\theta_1; z)]$

- A1. $\mu-$strongly convex. A2. $L$-jointly Lipschitz gradient.
- A3. ($\epsilon-$sensitivity) $W_1(\mathcal{D}(\theta), \mathcal{D}(\theta')) \le \epsilon\|\theta - \theta'\|$, $\forall\, \theta, \theta' \in \mathbb{R}^d$,
- A4. $\sigma$-perturbation with sampled gradient

$$\sup_{z\in\mathsf{Z}} \|\nabla\ell(\theta; z) - \nabla f(\theta; \theta_{PS})\| \le \sigma(1 + \|\theta - \theta_{PS}\|)$$

- A5. **Poisson equation**: $\exists\widehat{\nabla\ell} : \mathbb{R}^d \times \mathsf{Z} \to \mathbb{R}^d$ s.t. $\forall\, \theta, \theta' \in \mathbb{R}^d, z \in \mathsf{Z}$,

$$\nabla\ell(\theta'; z) - \mathbb{E}_{z'\sim\mathcal{D}(\theta)}[\nabla\ell(\theta'; z)] = \widehat{\nabla\ell}(\theta'; z) - \mathrm{P}_\theta\widehat{\nabla\ell}(\theta'; z).$$

- A6. $L_{PH}$-**smoothness** for Poisson equation

$$\sup_{z\in\mathsf{Z}} \|\mathrm{P}_\theta\widehat{\nabla\ell}(\theta; z) - \mathrm{P}_{\theta'}\widehat{\nabla\ell}(\theta'; z)\| \le L_{PH}\|\theta - \theta'\|, \ \forall\, \theta, \theta'.$$

- **Example**: For all $\theta$, uniformly geometrically ergodic MC ($\mathrm{P}_\theta$) (A5 ✓).
- **Example**: Agents' utility function $U(\cdot)$ is quadratic (A6 ✓ ).

**Theorem** Under A1-A6. Let $\epsilon < \frac{\mu}{L}$ and with non-increasing step sizes $\{\gamma_k\}_{k\ge1}$, there exists $\mathbb{C}$ where it holds

$$\mathbb{E}[\|\theta_k - \theta_{PS}\|^2] \le \underbrace{\prod_{i=1}^{k}\left(1 - \gamma_i\tfrac{\mu - L\epsilon}{2}\right)\|\theta_0 - \theta_{PS}\|^2}_{\text{Transient}} + \underbrace{\mathbb{C}\gamma_k}_{\text{Fluctuation}},$$

- Convergence region: $\epsilon < \frac{\mu}{L}$, Convergence rate $\mathcal{O}(1/k)$.
- (Non-convex $\ell$) Convergence to near-stationary point of $\mathcal{L}(\theta)$.

## Numerical Experiments

**Logistic Regression** —

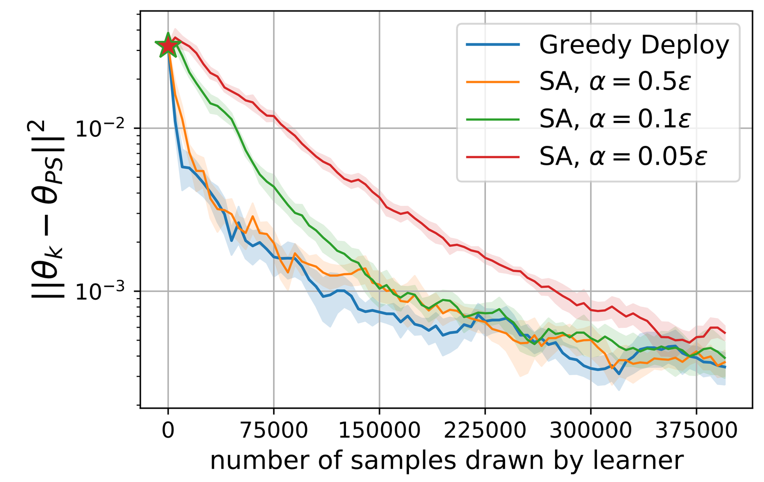
$$\ell(\theta; z) = \frac{\beta}{2}\|\theta\|^2 + \log(1 + \exp(\langle\theta, x\rangle)) - y\langle\theta, x\rangle$$

**Synthetic Data** for SVM

- **Goal**: examine the impact of different agents' response rate ($\alpha$) on convergence rate of SA.
- **Agent Response**: $\mathcal{D}(\theta)$ is obtained through evaluating the best quadratic response, i.e.

$$z_{k+1} \in \arg\max_{z'\in\mathsf{Z}} U(z'; \tilde{z}_{k+1} \sim \mathcal{D}_0),$$

where $U_q(z'; z, \theta) = \langle\theta, x'\rangle - \frac{\|x'-x\|^2}{2\epsilon}$

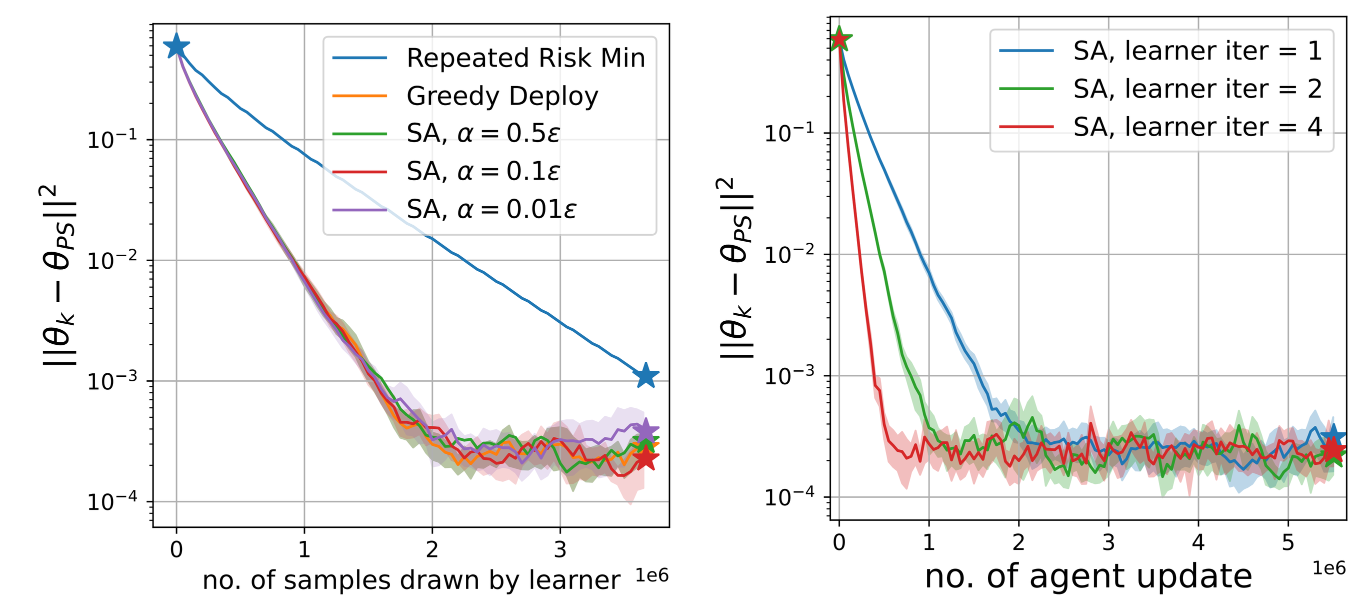



- $\alpha\downarrow$ leads to a slower Markov chain, the mixing time $\widehat{L}\uparrow$.

**Real Data** for strategic classification:

- **Agent** - Logistic response (no closed form solution)

$$U_{lg}(z'; z, \theta) = y\langle\theta, x'\rangle - \log(1 + \exp(\langle\theta, x'\rangle)) - \frac{\|x'-x\|^2}{2\epsilon}$$




- (Left) $\alpha\downarrow 0$, SA convergence speed $\downarrow$, mixing time $\uparrow$.
- (Right) no. learner's iteration $\uparrow \Rightarrow$ error decreases rate $\uparrow$.

**Reference**


- Perdomo, Juan, et al. *Performative prediction*, ICML 2020.
- Mendler-Dünner, et al. *Stochastic optimization for performative prediction* NeurIPS 2020.
- Brown, Gavin, et al. *Performative prediction in a stateful world*, arXiv:2011.03885 (2020).


Contact E-mail: {liqiang,htwai}@se.cuhk.edu.hk
Acknowledgement: CUHK Direct Grant #4055113.