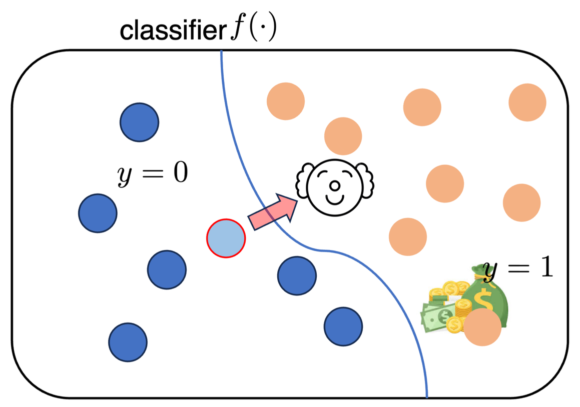
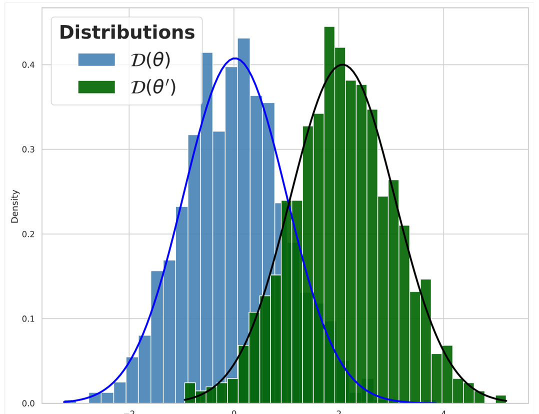
Stochastic Optimization Schemes for Performative Prediction with Nonconvex Loss

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

Performative Prediction

- **Motivation**: Learning in economic or societal environment is causative.
- **Example**: Hiring, Loan application.





- **Perf Pred**: *model to be trained can influence the outcome they aim to predict.*

Formulation

- *Performativity* modeled by **distribution shift** $\mathcal{D}(\boldsymbol{\theta})$.
- **Performative Risk Minimization**:

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

- But $\nabla V(\boldsymbol{\theta})$ is difficult to estimate $\Rightarrow$

SGD-Greedy Deploy (SGD-GD):

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma\nabla\ell(\boldsymbol{\theta}_t; Z_{t+1}),\ Z_{t+1}\sim\mathcal{D}(\boldsymbol{\theta}_t)$$

- Leads to a **non-gradient dynamics**.
- **Fact** [Mendler-Dünner, 2020]: If $\ell(\boldsymbol{\theta}; Z)$ = str. cvx & mild dist. shift, then SGD-GD $\to$ 'performative stable' (PS) sol:

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

- **Limitation**: requires str. cvx $\ell(\cdot; z)$.

δ-Stationary Perf. Stable sol.

- **Def.** $\boldsymbol{\theta}^\star\in\mathbb{R}^d$ is an δ-SPS solution if:

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

- If $\ell(\boldsymbol{\theta}; z)$ is str. cvx, then (0-)SPS = PS.

Key Takeaways

- **(A)** SGD w/ greedy deployment finds an $\mathcal{O}(\epsilon)$-biased SPS sol.
- **(B)** Bias level reduced to $\mathcal{O}(\epsilon^2)$ with exact gradient.
- **(C)** SGD w/ lazy deployment finds bias-free SPS sol if ep. length $\to\infty$.
- **Idea**: time varying Lyapunov function for non-gradient dynamics.

Main Results

- Set $J(\boldsymbol{\theta}_1; \boldsymbol{\theta}_2) := \mathbb{E}_{Z\sim\mathcal{D}(\boldsymbol{\theta}_2)}[\ell(\boldsymbol{\theta}; Z)]$, partial gradient $\nabla_1 J(\boldsymbol{\theta}_1; \boldsymbol{\theta}_2) := \mathbb{E}_{Z\sim\mathcal{D}(\boldsymbol{\theta}_2)}[\nabla\ell(\boldsymbol{\theta}; Z)]$.

- **A1. (Smoothness)** $\|\nabla\ell(\boldsymbol{\theta}; z) - \nabla\ell(\boldsymbol{\theta}'; z)\| \le L\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|, \forall\boldsymbol{\theta},\boldsymbol{\theta}'\in\mathbb{R}^d$.
- **A2. (Variance)** $\mathbb{E}_{Z\sim\mathcal{D}(\boldsymbol{\theta}_2)}\left[\|\nabla\ell(\boldsymbol{\theta}_1; Z) - \nabla_1 J(\boldsymbol{\theta}_1; \boldsymbol{\theta}_2)\|^2\right] \le \sigma_0^2 + \sigma_1^2\|\nabla J(\boldsymbol{\theta}_1; \boldsymbol{\theta}_2)\|^2$.

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

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

- Note that **C1** = stronger than **W1**, but **C2** = weaker than **W2**.

Theorem 1: Under **A1-2**, (**C1** & **C2**) or (**W1** & **W2**). It holds

$$\mathbb{E}\left[\|\nabla_1 J(\boldsymbol{\theta}_{\mathsf{T}}; \boldsymbol{\theta}_{\mathsf{T}})\|^2\right] \lesssim 1/\sqrt{T} + \underbrace{\tilde{L}\epsilon\,(\sigma_0 + (1+\sigma_1^2)\tilde{L}\epsilon)}_{\mathcal{O}(\epsilon\sigma_0+\epsilon^2)-\textbf{bias}}.$$

- Biased-SPS Sol.: $\mathcal{O}(\epsilon)$ for noisy SGD, $\mathcal{O}(\epsilon^2)$ for noiseless SGD.
- **Proof Idea**: study a *descent lemma* for $J(\boldsymbol{\theta}_{t+1}; \boldsymbol{\theta}_t) - J(\boldsymbol{\theta}_t; \boldsymbol{\theta}_t)$, then bound the distance for $|J(\boldsymbol{\theta}_{t+1}; \boldsymbol{\theta}_t) - J(\boldsymbol{\theta}_{t+1}; \boldsymbol{\theta}_{t+1})|$.

SGD-Lazy Deployment: let $K\ge 1$ be the epoch length

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

Theorem 2: Same as **Theorem 1** + bounded gradient. It holds

$$\mathbb{E}\left[\|\nabla_1 J(\boldsymbol{\theta}_{\mathsf{T}}; \boldsymbol{\theta}_{\mathsf{T}})\|^2\right] \lesssim \frac{1}{\sqrt{T}} + \frac{L\sigma_0^2}{K\sqrt{T}} + \frac{\tilde{L}\epsilon}{K}\left(\sqrt{K}\sigma_0 + \sqrt{(K+\sigma_1^2)\tilde{L}\epsilon}\right).$$

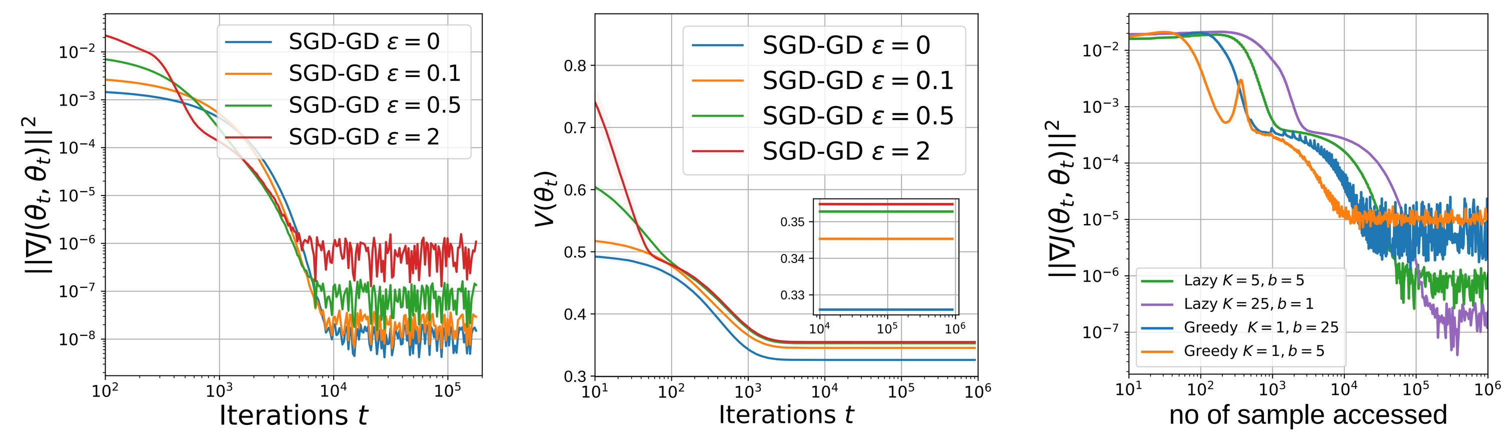
- With $K\uparrow\infty$, lazy deployment $\approx$ repeated risk minimization.
- Finds a *bias-free SPS solution* when $T, K\uparrow\infty$.

Synthetic Data with Linear Model

- **Setup**: sigmoid loss (smooth but non-convex).

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

- **Data & Dist. Shift**: $\mathcal{D}^o \equiv \{(x_i, y_i)\}_{i=1}^m$, $x_i\sim\mathcal{U}[-1,1]^d$, $y_i = \mathsf{sgn}(\langle x_i, \boldsymbol{\theta}^o\rangle)\in\{\pm 1\}$, $\mathcal{D}(\boldsymbol{\theta}) = \{(x_i - \epsilon_L\boldsymbol{\theta}, y_i)\}_{i=1}^m$.



- (Left) SGD-GD converges to a biased-SPS solution. $\epsilon_L\uparrow\to$ bias $\uparrow$. $\to$ **Theorem 1** ✓
- (Middle) Performative risk $V(\boldsymbol{\theta}_t)$ vs Iterations t. $\epsilon_L\uparrow$ leads to higher bias.
- (Right) Set stepsize $\gamma = 1/(K\sqrt{T})$. $K\uparrow$ leads to lower bias. $\to$ **Theorem 2** ✓

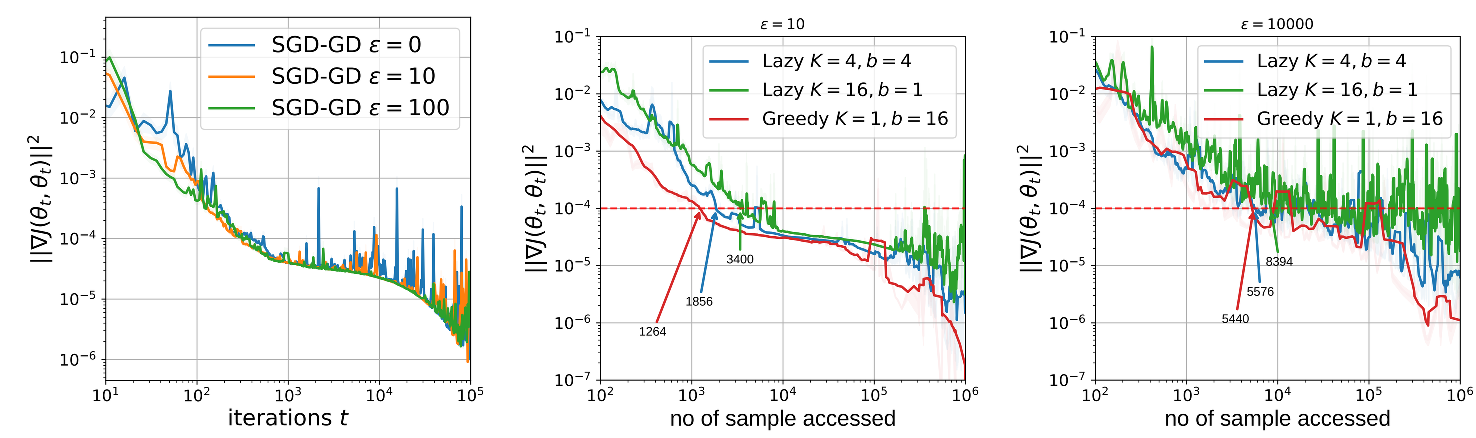
Spam Detection with Neural Network

- **Data**: Spambase [Hopkins et al. 1999]. $m = 4601$, $d = 48$.
- **NN Classifier** $f_{\boldsymbol{\theta}}(x)$: three fully-connected layers with $\tanh$ activation and a sigmoid output layer.
- **Distribution Shift**: $z\equiv(x,\bar{y})\sim\mathcal{D}(\boldsymbol{\theta})$.

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

We approximate $x\approx\bar{x} - \epsilon_{\mathsf{NN}}\nabla_x f_{\boldsymbol{\theta}}(\bar{x})$, $\epsilon\propto\epsilon_{\mathsf{NN}}$.

- **Param.**: $\gamma_{\mathsf{Greedy}} = 200/\sqrt{T}$, $\gamma_{\mathsf{Lazy}} = 200/(K\sqrt{T})$.



- Lazy deploy performs better than greedy as $\epsilon_{\mathsf{NN}}\uparrow$. When $\epsilon_{\mathsf{NN}}: 10\mapsto 10^5$, no. sample for three algo: $\times 4$, $\times 3$, $\times 2.4$.

References

- Perdomo, Juan, et al. *Performative prediction*, ICML 2020.
- Mendler-Dünner, et al. *Stochastic optimization for performative prediction* NeurIPS 2020.

Contact E-mail: {liqiang,htwai}@se.cuhk.edu.hk

Acknowledgement: CUHK Direct Grant #4055208.