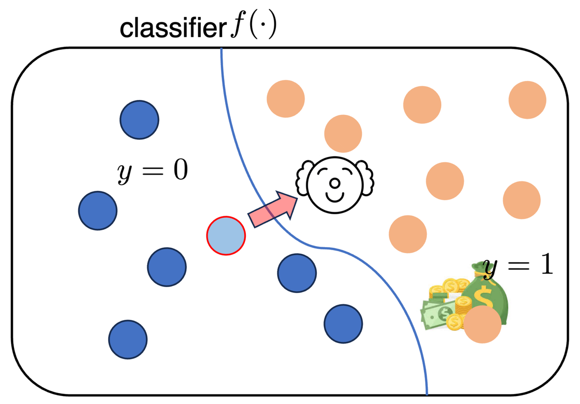
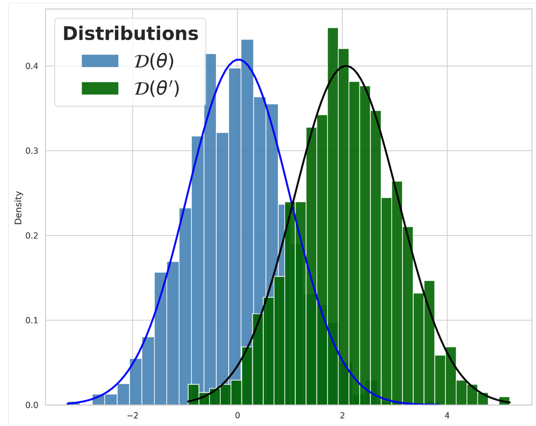
Clipped SGD Algorithms for Performative Prediction: Tight Bounds for Clipping Bias and Remedies

Qiang Li* Michal Yemini† Hoi-To Wai* * Dept. of SEEM, CUHK, † Fac. of Engineering, Bar-Ilan University.

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** $\mathcal{D}(\boldsymbol{\theta})$.
- Let $\ell(\boldsymbol{\theta}; Z)$ be the loss function to be minimized,

SGD-Greedy Deploy (SGD-GD):

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

- Interaction between *learner* and *data*.
- **Risk**: model inversion attack [Ghosh et al., 2009] exposes sensitive user data using just the training history of SGD.

Convergence Metrics (str/non cvx)

- **Def.** Performative stable (PS) solution:

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

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

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

- If $\ell(\boldsymbol{\theta}; z)$ is strongly convex, then (0-)SPS $\Longleftrightarrow$ PS solution.

Two Clipping Algorithms

- Draw sample $Z_{t+1} \sim \mathcal{D}(\boldsymbol{\theta}_t)$, injected noise $\zeta_{t+1} \sim \mathcal{N}(0, \sigma_{\text{DP}}^2 \boldsymbol{I})$.

Projected Clipped SGD (PCSGD):

$$\boldsymbol{\theta}_{t+1} = \mathcal{P}_{\mathcal{X}}\left(\boldsymbol{\theta}_t - \gamma_{t+1} \text{clip}_c\left[\nabla \ell(\boldsymbol{\theta}_t; Z_{t+1}] + \zeta_{t+1}\right),\right.$$

where $\mathcal{P}_{\mathcal{X}}(\cdot)$ is project operator, clipping operator is

$$\text{clip}_c(\boldsymbol{g}) : \boldsymbol{g} \in \mathbb{R}^d \mapsto \min\left\{1, \frac{c}{\|\boldsymbol{g}\|_2}\right\} \boldsymbol{g} \rightarrow \text{reduce gradient exposure}$$

DiceSGD [Zhang et al., 2024]: $v_{t+1} = \text{clip}_{C_1}(\nabla \ell(\boldsymbol{\theta}_t; Z_{t+1})) + \text{clip}_{C_2}(e_t)$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_{t+1}(v_{t+1} + \zeta_{t+1}), \quad e_{t+1} = e_t + \nabla \ell(\boldsymbol{\theta}_t; Z_{t+1}) - v_{t+1}$$

Main Results

- Set $f(\boldsymbol{\theta}_1; \boldsymbol{\theta}_2) := \mathbb{E}_{Z \sim \mathcal{D}(\boldsymbol{\theta}_2)}[\ell(\boldsymbol{\theta}; Z)]$, partial gradient $\nabla f(\boldsymbol{\theta}_1; \boldsymbol{\theta}_2) := \mathbb{E}_{Z \sim \mathcal{D}(\boldsymbol{\theta}_2)}[\nabla \ell(\boldsymbol{\theta}; Z)]$.
  - **A1**. **(Smoothness)** $\|\nabla \ell(\boldsymbol{\theta}; z) - \nabla \ell(\boldsymbol{\theta}'; z')\| \leq L(\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| + \|z - z'\|)$.
  - **A2**. **(Variance)** $\mathbb{E}_{Z \sim \mathcal{D}(\boldsymbol{\theta}_2)}\left[\|\nabla \ell(\boldsymbol{\theta}_1; Z) - \nabla f(\boldsymbol{\theta}_1; \boldsymbol{\theta}_2)\|^2\right] \leq \sigma_0^2 + \sigma_1^2 \|\nabla f(\boldsymbol{\theta}_1; \boldsymbol{\theta}_2)\|^2$.
  - **A3**. **(Bounded Gradient)** There exists $G \geq 0$ s.t. $\sup_{\boldsymbol{\theta} \in \mathcal{X}, z \in \mathsf{Z}} \|\nabla \ell(\boldsymbol{\theta}; z)\| \leq G$.
  - **A4**. **(Wasserstein sensitivity)** $\mathcal{W}_1(\mathcal{D}(\boldsymbol{\theta}), \mathcal{D}(\boldsymbol{\theta}')) \leq \beta \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$.

  - **C1**: **(TV sensitivity)**: Total variation distance $d_{\text{TV}}(\mathcal{D}(\boldsymbol{\theta}), \mathcal{D}(\boldsymbol{\theta}')) \leq \beta \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$.
  - **C2**: **(Bounded loss)**: There exists $\ell_{\max} \geq 0$ s.t. $\sup_{\boldsymbol{\theta} \in \mathbb{R}^d, z \in \mathsf{Z}} |\ell(\boldsymbol{\theta}; z)| \leq \ell_{\max}$.

Theorem 1: Under **A1,3,4**. Suppose $\beta < \frac{\mu}{L}$, $f(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}})$ is strongly convex *w.r.t.* $\boldsymbol{\theta}$ and denote $\widetilde{\mu} := \mu - L\beta$, then the iterates of **PCSGD** hold that

$$\mathbb{E}[\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_{PS}\|^2] \leq \prod_{i=1}^{t+1}(1 - \widetilde{\mu}\gamma_i)\|\hat{\boldsymbol{\theta}}_0\|^2 + \frac{2(c^2 + G^2)}{\widetilde{\mu}}\gamma_{t+1} + \frac{8(\max\{G - c, 0\})^2}{\widetilde{\mu}^2},$$

Theorem 2: bias order is **tight**: Bias $= \Theta(1/(\mu - L\beta)^2)$, which increases as $\beta \uparrow \frac{\mu}{L}$.

Theorem 3: Suppose $f(\cdot; \boldsymbol{\theta})$ is non-cvx. Under **A1,2,3**, **C1,2**. **PCSGD** holds that

$$\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_{\mathsf{T}}; \boldsymbol{\theta}_{\mathsf{T}})\|^2] \lesssim \frac{1}{\sqrt{T}} + \mathcal{O}(\ell_{max}\beta + \max\{G - c, 0\}^2).$$

- **Idea**: time varying Lyapunov function for **non-gradient and non-smooth dynamics**.
- **Cor. 1** (Privacy Guarantee) **PCSGD** is (ε, δ)-DP if we let $\sigma_{\text{DP}} \geq c\sqrt{T \log(1/\delta)}/(m\varepsilon)$.
- **Cor. 2** (Optimal Constant Stepsize) To reduce bias, we set $\gamma^\star = \widetilde{\mathcal{O}}((\widetilde{\mu}T)^{-1})$.
- **Cor. 3** (Optimal Clipping Threshold) To achieve opt. asymptotic ub, $c^\star = \frac{2Gm^2\varepsilon^2}{d\log(1/\delta) + 2m^2\varepsilon^2}$.

Reducing Clipping Bias in Clipped SGD

- **DiceSGD**: error feedback mechanism is effective in removing the asymptotic bias. Since the fixed point $(\bar{e}; \bar{\boldsymbol{\theta}})$ satisfies

$$-\text{clip}_{C_2}(\bar{e}) = \mathbb{E}_{Z \sim \mathcal{D}(\bar{\boldsymbol{\theta}})}[\text{clip}_{C_1}(\nabla \ell(\bar{\boldsymbol{\theta}}; Z))]$$
$$\nabla f(\bar{\boldsymbol{\theta}}; \bar{\boldsymbol{\theta}}) - \text{clip}_{C_2}(\bar{e}) = \mathbb{E}_{Z \sim \mathcal{D}(\bar{\boldsymbol{\theta}})}[\text{clip}_{C_1}(\nabla \ell(\bar{\boldsymbol{\theta}}; Z))]$$

- If $C_2 \geq C_1$, fixed point $(\bar{e}; \bar{\boldsymbol{\theta}})$ satisfies $\nabla f(\bar{\boldsymbol{\theta}}; \bar{\boldsymbol{\theta}}) = \mathbf{0}$.

Theorem 4: Suppose that $f(\cdot; \boldsymbol{\theta})$ is strongly convex and $\beta < \frac{\mu}{L}$. Under **A1,2,4** and mild assumptions, **DiceSGD** holds

$$\mathbb{E}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{PS}\|^2 \leq 2\mathbb{E}\|\tilde{\boldsymbol{\theta}}_t\|^2 + 2\gamma_t^2 \mathbb{E}\|e_t\|^2 = \mathcal{O}(1/t),$$

where $\tilde{\boldsymbol{\theta}}_t := \boldsymbol{\theta}_t - \gamma_t e_t$.

Theorem 5: Suppose that $f(\cdot; \boldsymbol{\theta})$ is non-convex. Under **A1,2**, **C1,2** and mild assumptions, **DiceSGD** holds

$$\min_{t=0,\ldots,T-1} \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t; \boldsymbol{\theta}_t)\|^2] = \mathcal{O}\left(1/\sqrt{T} + \mathsf{b}\beta\right),$$

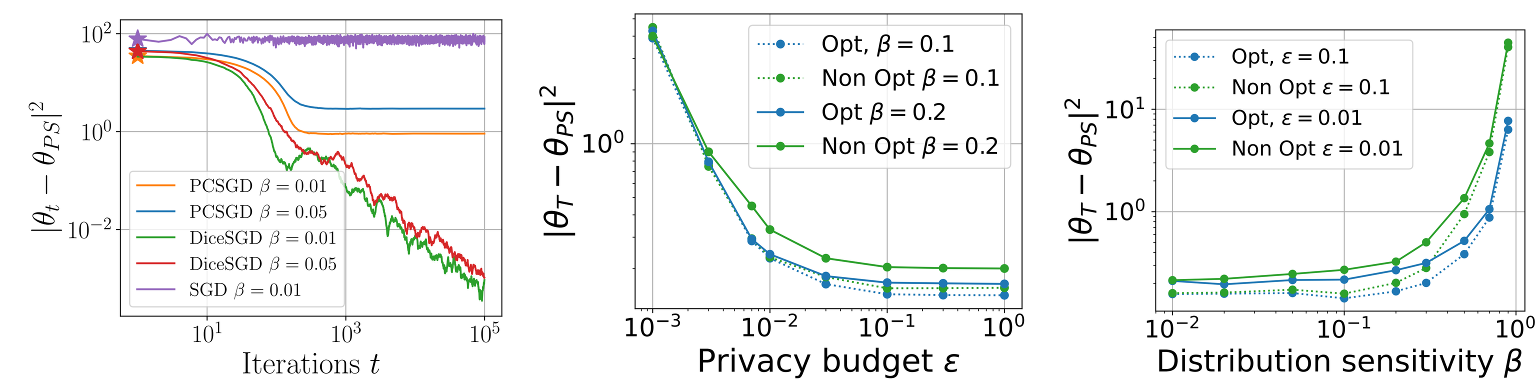
where $\mathsf{b} = \mathcal{O}(\ell_{\max}((C_1 + C_2) + \sqrt{d}\sigma_{\text{DP}}))$.

Quadratic Min. with Synthetic Data

- Consider a scalar performative risk problem,

$$\min_{\boldsymbol{\theta} \in \mathcal{X}} \mathbb{E}_{z \sim \mathcal{D}(\boldsymbol{\theta})}[(\boldsymbol{\theta} + az)^2/2],$$

- **Data**: $\mathcal{D}(\boldsymbol{\theta}) = \text{Unif}\left(\{b\tilde{Z}_i - \beta\boldsymbol{\theta}\}_{i=1}^m\right)$, where $\tilde{Z}_i \sim \mathcal{B}(p)$ is Bernoulli. $\boldsymbol{\theta}_{PS} = -\bar{p}a/(1 - a\beta)$, where $\bar{p}$ is sampel mean.



- **(Left)** SGD w/ DP noise can not converge. PCSGD converge to $\boldsymbol{\theta}_{PS}$ with bias which increase as $\beta \uparrow$ **[Thm1&2 ✓]**
- DiceSGD finds bias-free sol. at rate of $\mathcal{O}(1/t)$ **[Thm4✓]**
- **(Middle & Right)** set opt step size $\gamma^\star$ adapted to dist. shift achieves smaller bias. **[Cor. 1 ✓]** As privacy budget $\varepsilon \downarrow$, or sensitivity $\beta \uparrow \frac{\mu}{L}$, the bias of **PCSGD** $\uparrow$ **[Cor. 2 ✓]**

References

- Perdomo, Juan, et al. *Performative prediction*, ICML 2020.
- Zhang et al., *Differentially private sgd without clipping bias*, ICML, 2024.
- Gosh et al., *Universally utility-max. privacy mechanisms*, ACM, 2009.
- Li and Wai, *Performative Prediction with NCVX Loss*, NeurIPS 2024.

Contact E-mail: {liqiang,htwai}@se.cuhk.edu.hk, michal.yemini@biu.ac.il

Acknowledgement: Shun Hing Institute of Advanced Engineering, CUHK, Project #MMT-p5-23