Privacy-Efficacy Tradeoff of Clipped SGD with Decision-dependent Data

Qiang Li, Michal Yemini, Hoi-To Wai

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

July 10, 2024

ICML2024 Workshop: Humans, Algorithmic Decision-Making and Society: Modeling Interactions and Impact

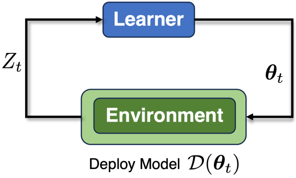
Privacy Concerns in Model Training

- The training of prediction models hinges on the use of **private and sensitive** user data such as credit history and customer's identity.
- **Risk**: many attack techniques can expose sensitive user data using just the training history of stochastic gradient descent (SGD) algorithms, such as,
  - Membership inference attack
  - Feature inference attack
  - Model extraction attack...

  launched by curious observers or adversary customers, [Ghosh et al., 2009, Bassily et al., 2014].

Real Example

- **Example**: online platforms (learner) collect data to train their bot detection models. Conversely, advertisers use bots to automate advertising campaigns such as clicking on ads and visiting websites.
- **Strategic Response** of advertisers: to bypass bot-detecting model, the advertiser trains a bot model, allowing them appear as human users, and uses it to *invert the predictions* of the bot detection model.
- **Distribution Shift**: user reacts to the changing models, also known as **performative prediction problem**.



Performative Prediction

- Performative Prediction refers to stochastic optimization problem based on **dynamical** data distribution.

$$\min_{\boldsymbol{\theta}\in\mathcal{X}} \mathbb{E}_{Z\sim\mathcal{D}(\boldsymbol{\theta})}[\ell(\boldsymbol{\theta};Z)], \tag{1}$$

  where $\mathcal{X}$ is the feasible region, $\mathcal{D}(\boldsymbol{\theta})$ is shift dist. induced by $\boldsymbol{\theta}$.
- **Supervised Learning vs Perf. Pred.**: $\mathcal{D}$ and $\mathcal{D}(\boldsymbol{\theta})$.
- **Related topics**: game theory and Stackberg games.
- **Performative stable solution**:

$$\boldsymbol{\theta}_{PS} = \arg\min_{\boldsymbol{\theta}\in\mathcal{X}} \ \mathbb{E}_{Z\sim\mathcal{D}(\boldsymbol{\theta}_{PS})}[\ell(\boldsymbol{\theta};Z)]. \tag{2}$$

  i.e., there is no incident of gradient at $\boldsymbol{\theta}_{PS}$,

$$\left\|\mathbb{E}_{Z\sim\mathcal{D}(\boldsymbol{\theta}_{PS})}[\nabla\ell(\boldsymbol{\theta}_{PS};Z)]\right\| = 0$$

Privacy Preserving ...

- **Privacy-preserving algorithm**: clipped SGD algorithm [Abadi et al., 2016] is designed to address above challenge.

Projected clipped SGD algorithm

$$\boldsymbol{\theta}_{t+1} = \mathcal{P}_{\mathcal{X}}\left(\boldsymbol{\theta}_t - \gamma_{t+1}\mathsf{clip}_c\left(\text{stoc. grad}\right) + \zeta_{t+1}\right)$$

where $\mathcal{P}(\cdot)$ is projection operator, $\zeta_{t+1} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathsf{DP}}^2 \boldsymbol{I})$ is noise.

- **Clipping operator**: designed to reduce grad. exposure [Pascanu et al., 2013],

$$\mathsf{clip}_c(\boldsymbol{g}) : \boldsymbol{g} \in \mathbb{R}^d \mapsto \min\left\{1, \frac{c}{\|\boldsymbol{g}\|_2}\right\}\boldsymbol{g}, \tag{3}$$

where c is clipping threshold.

Research Gap

- **Related works**: [Koloskova et al., 2023] shows that clipped SGD may only converge to a near critical points solution in stochastic setting, due to the unavoidable bias introduced by clip operator.
- Most studies on clipped SGD algo. is in the **absence of performativity**, i.e., ignoring the effects of the decision-dependent distribution.

Question: *What effect does performativity have on bias and convergence of clipped SGD algorithms?*

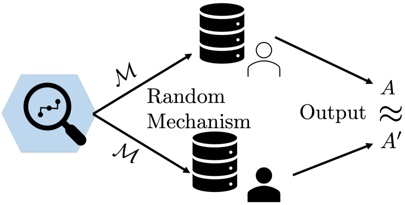
Our Answer: Projected Clipping SGD (PCSGD) algorithm converges to a biased solution in expectation. We found bias amplification effect, i.e., bias $\propto \mathcal{O}(1/\text{dist. shift sensitivity})$.

Preliminaries

- $(\varepsilon, \delta)-$**differential privacy** [Dwork and Roth, 2014]: $\mathcal{M} : \mathcal{D} \mapsto \mathcal{R}$ is a randomized mechanism. For adjacent inputs $\mathsf{D}, \mathsf{D}' \in \mathcal{D}$, which differs by only 1 different sample, it holds that

$$\Pr[\mathcal{M}(\mathsf{D}) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(\mathsf{D}') \in S] + \delta. \tag{4}$$

- (ε, δ) is called privacy budget, δ is probability of information leakage.
- When $\varepsilon \approx 0$, then $\Pr[\mathcal{M}(\mathsf{D}) \in S] \approx \Pr[\mathcal{M}(\mathsf{D}') \in S]$, i.e., output of $\mathcal{M}(\cdot)$ does not vary whether a record is present or absent from the system.



Clipped SGD Algorithms

To ensure **privacy**, we study the following **PCSGD** scheme:

$$\boldsymbol{\theta}_{t+1} = \mathcal{P}_{\mathcal{X}}\big(\boldsymbol{\theta}_t - \gamma_{t+1}(\mathsf{clip}_c(\nabla \ell(\boldsymbol{\theta}_t; Z_{t+1})) + \zeta_{t+1})\big), \qquad (5)$$

- Fixed point $\boldsymbol{\theta}_\infty$ of **PCSGD** satisfies

$$\mathbb{E}_{Z\sim\mathcal{D}(\boldsymbol{\theta})}[\mathsf{clip}_c(\nabla \ell(\boldsymbol{\theta}; Z))] = \mathbf{0}.$$

- In general, $\boldsymbol{\theta}_\infty$ will leads to $\|\mathbb{E}[\nabla \ell(\boldsymbol{\theta}_\infty; Z)]\| \neq 0$ and vice versa due to clipping operator.
- **Challenge**: clipping operator is non-smooth and leads to

$$\mathbb{E}_{Z\sim\mathcal{D}(\boldsymbol{\theta})}\mathsf{clip}_c(\nabla \ell(\boldsymbol{\theta}; Z)) \neq \mathbb{E}_{Z\sim\mathcal{D}(\boldsymbol{\theta})}(\nabla \ell(\boldsymbol{\theta}; Z))$$

  In other words, the stochastic gradient is not unbiased estimation of its expectation. Additionally, performativity will even exacerbate this issue.
- **Greedy deployment** sampling scheme: $Z_{t+1} \sim \mathcal{D}(\boldsymbol{\theta}_t)$.

Assumptions & Notations

Define the shorthand notations:

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

- A1: μ-strongly convex w.r.t. $\boldsymbol{\theta}$:

$$f(\boldsymbol{\theta}'; \bar{\boldsymbol{\theta}}) \geq f(\boldsymbol{\theta}; \bar{\boldsymbol{\theta}}) + \langle \nabla f(\boldsymbol{\theta}; \bar{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle + (\mu/2) \left\| \boldsymbol{\theta}' - \boldsymbol{\theta} \right\|^2 .$$

- A2: Maps $\nabla f(\cdot; \bar{\boldsymbol{\theta}})$ and $\nabla \ell(\bar{\boldsymbol{\theta}}; \cdot)$ are L-Lipschitz:

$$\|\nabla f(\boldsymbol{\theta}_1; \bar{\boldsymbol{\theta}}) - \nabla f(\boldsymbol{\theta}_2; \bar{\boldsymbol{\theta}})\| \leq L\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|, \ \forall \ \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{X},$$
$$\|\nabla \ell(\bar{\boldsymbol{\theta}}; z_1) - \nabla \ell(\bar{\boldsymbol{\theta}}; z_2)\| \leq L \left\| z_1 - z_2 \right\|, \ \forall z_1, z_2 \in \mathsf{Z}.$$

- A3: Wasserstein-1 distance:

$$W_1(\mathcal{D}(\boldsymbol{\theta}), \mathcal{D}(\boldsymbol{\theta}')) \leq \beta\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.$$

- A4 : Uniform bound: There exists $G \geq 0$ such that

$$\sup_{\boldsymbol{\theta}\in\mathcal{X}, z\in\mathsf{Z}} \|\nabla \ell(\boldsymbol{\theta}; z)\| \leq G$$

Note that A4 assumes bounded gradient but only on a compact set $\mathcal{X}$ which is reasonable, see [Zhang et al., 2024].

Main Results (I)

Define the following constants: $c_1 := 2(c^2 + G^2) + d\sigma_{\mathsf{DP}}^2$, $\mathcal{C}_1 := (\max\{G - c, 0\})^2$, $\tilde{\mu} := \mu - L\beta$.

Theorem 1: Upper bound

Under A1-4. Suppose that $\beta < \frac{\mu}{L}$, the step sizes $\{\gamma_t\}_{t\geq 1}$ are non-increasing and are sufficient small. Then, for any $t \geq 1$,

$$\mathbb{E}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_{PS}\|^2 \leq \prod_{i=1}^{t+1}(1 - \tilde{\mu}\gamma_i)\,\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_{PS}\|^2 + \frac{2c_1}{\tilde{\mu}}\gamma_{t+1} + \underbrace{\frac{8\mathcal{C}_1}{\tilde{\mu}^2}}_{\text{bias}},$$

- It indicates an asymptotic clipping bias of **PCSGD** and coincides with the observation in [Koloskova et al., 2023] for non-decision dependent distribution.
- When $c \geq G$, then $\mathcal{C}_1 = 0$ and the bias vanishes. Our convergence rate $\mathcal{O}(\gamma_t)$ coincides with [Drusvyatskiy and Xiao, 2023].

Main Results (II)

Theorem 2: Lower bound

For any $c \in (0, G)$, there exist $\ell(\boldsymbol{\theta}; Z)$ and $\mathcal{D}(\boldsymbol{\theta})$ satisfying A1-4, such that for all fixed-points of **PCSGD** $\boldsymbol{\theta}_\infty$ satisfying $\mathbb{E}_{Z \sim \mathcal{D}(\boldsymbol{\theta}_\infty)}[\text{clip}_c(\nabla \ell(\boldsymbol{\theta}_\infty; Z))] = \mathbf{0}$, it holds that

$$\|\boldsymbol{\theta}_\infty - \boldsymbol{\theta}_{PS}\|^2 = \Omega\left(1/(\mu - L\beta)^2\right). \tag{6}$$

- Provided that $\beta < \frac{\mu}{L}$, Theorems 1 and 2 show that **PCSGD** admits an unavoidable bias of $\Theta(1/(\mu - L\beta)^2)$.

Corollary 1: Differential Privacy Guarantee [Abadi et al., 2016]

For any $\varepsilon \leq T/m^2$, $\delta \in (0, 1)$, and $c > 0$, the **PCSGD** with greedy deployment is (ε, δ)-DP after T iterations if we let

$$\sigma_{\text{DP}} \geq c\sqrt{T \log(1/\delta)}/(m\varepsilon).$$

- Assume that $G > c$ and a constant step size is used in **PCSGD**. To achieve *minimum bias*, we can compute $\gamma^\star = \mathcal{O}(1/(\tilde{\mu}T))$.

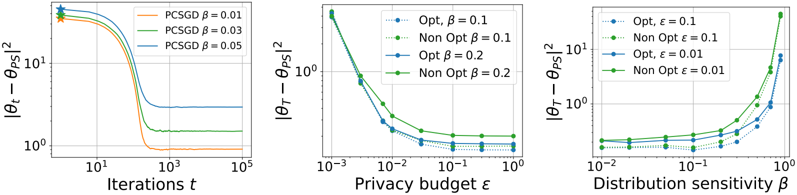
Numerical Simulation

Quadratic Minimization

- We consider a scalar performative risk problem
$$\min_{\boldsymbol{\theta}\in\mathcal{X}} \mathbb{E}_{z\sim\mathcal{D}(\boldsymbol{\theta})}[(\boldsymbol{\theta}+az)^2/2]$$
Distribution is set as $\mathcal{D}(\boldsymbol{\theta}) = \{b\tilde{Z}_i - \beta\boldsymbol{\theta}\}_{i=1}^m$ where $\tilde{Z}_i \sim \mathcal{B}(p)$ is Bernoulli and $a > 0, b > 0$, $p < 1/2$.
- Performative stable solution: $\boldsymbol{\theta}_{PS} = \frac{-\bar{p}a}{1-a\beta}$, where $\bar{p} = \frac{1}{m}\sum_{i=1}^m \tilde{Z}_i$.

- **Settings:** $p = 0.1, \varepsilon = 0.1, \delta = 1/m, \beta \in \{0.01, 0.05\}, a = 10, b = 1, c = C_1 = C_2 = 1$, $m = 10^5$. The step size is $\gamma_t = \frac{10}{100+t}$ with the initialization $\boldsymbol{\theta}_0 = 5$.

Simulation (Cont'd)



- **Verifying Theorem 1 & 2**: In left fig, **PCSGD** can not converge to $\boldsymbol{\theta}_{PS}$ due to bias which increases as $\beta \uparrow$.
- **Trade off between bias and privacy budget**: From middle fig., $\epsilon \uparrow$ will leads to bias $\downarrow$. Also, optimal step size $\gamma^\star$ can achieve min bias, non-opt step size $\gamma = \frac{\log(1/\Delta(\mu))}{\mu T}$ has larger bias.
- **Trade off between bias and dist. shift**: From right fig., as the sensitivity of distribution shift increases $\beta \uparrow \frac{\mu}{L}$, the bias of **PCSGD** increases.

Conclusion

- We consider the privacy performative prediction problem and demonstrate that **PCSGD** can converge to a near-performative stable solution.
- **Key Observation**: The method exploits the bias amplification phenomenon caused by distribution shift.
- **Limitation/Ongoing Work**: Efforts are ongoing to reduce bias to an approximation of zero.

Questions & Comments?

References I

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016).
Deep learning with differential privacy.
In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Bassily, R., Smith, A., and Thakurta, A. (2014).
Private empirical risk minimization: Efficient algorithms and tight error bounds.
In *2014 IEEE 55th annual symposium on foundations of computer science*, pages 464–473. IEEE.

Drusvyatskiy, D. and Xiao, L. (2023).
Stochastic optimization with decision-dependent distributions.
Mathematics of Operations Research, 48(2):954–998.

Dwork, C. and Roth, A. (2014).
The algorithmic foundations of differential privacy.
Foundations and Trends® in Theoretical Computer Science, 9(3–4):211–407.

Ghosh, A., Roughgarden, T., and Sundararajan, M. (2009).
Universally utility-maximizing privacy mechanisms.
In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 351–360.

References II

Koloskova, A., Hendrikx, H., and Stich, S. U. (2023).
Revisiting gradient clipping: Stochastic bias and tight convergence guarantees.
arXiv preprint arXiv:2305.01588.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013).
On the difficulty of training recurrent neural networks.
In *International conference on machine learning*, pages 1310–1318. Pmlr.

Zhang, X., Bu, Z., Wu, Z. S., and Hong, M. (2024).
Differentially private sgd without clipping bias: An error-feedback approach.
In *ICLR*.