On the Sample Complexity of Differentially Private Policy Optimization

Yi He Xingyu Zhou

Wayne State University

Motivation and Key Takeaways

Motivation

As PO becomes increasingly prevalent in real-world applications, privacy concerns are emerging as a critical challenge. (e.g., patient interactions in personalized medical care, user prompts in large language models (LLMs))

Key question

What is the sample complexity cost induced by differential privacy in policy optimization?

Main contributions

1. **PO-specific DP definition**: We propose a DP notion tailored for PO, accounting for unique learning dynamics and privacy units.
2. **Unified meta-algorithm**: Enables private PG, NPG, and REBEL; reduces PO to private regression in some cases.
3. **Sample Complexity**: Our theoretical results demonstrate that privacy costs can often manifest as lower-order terms in the sample complexity.

Key Takeaways

1. **Privacy can be achieved with minimal statistical cost**: leading terms match non-private bounds(such that Yuan et al.[4]).
2. **Specific problem structures matters**: often lead to better results, both statistically and computationally.

Differential Privacy in Policy Optimization

Definition 1 : DP in PO

- Consider any policy optimization algorithm $\mathcal{M}$ interacting with a set D of N "users" and $\mathcal{M}(D)$ being the final output policy. We say $\mathcal{M}$ is (ε, δ)-DP if for adjacent datasets D, D' differing by one "user", and $\forall S \subseteq \text{Range}(\mathcal{M})$:

$$\mathbb{P}[\mathcal{M}(D) \in S] \leqslant e^{\varepsilon} \cdot \mathbb{P}[\mathcal{M}(D') \in S] + \delta.$$

- **Remark:** The standard DP definition assumes a fixed dataset of i.i.d. samples and protects the privacy of individual data records, making it suitable for supervised learning. In contrast, policy optimization (PO) involves dynamically collected data through on-policy interactions, where changing one sample can influence future data due to policy shifts, in that case, our DP in PO redefines the privacy unit as a "user" (e.g., a patient or prompt).

A Meta Algorithm for Private PO

Algorithm 1: A Meta Algorithm

`// Input: reward function` r`, learning rate` η`, batch size` m`, policy class` π_θ`, base policy` μ`, and a PrivUpdate oracle`

1. **Initialize:** $\theta_1 = 0$
2. **For** $t = 1$ to T:
   - Collect a fresh dataset $\bar{D}_t = \{(x_i, y_i, y_i')\}_{i=1}^m$ where:
   $$x_i \sim \rho, \quad y_i \sim \mu(\cdot|x_i), \quad y_i' \sim \pi_{\theta_t}(\cdot|x_i)$$
   - For all $i \in [m]$, let $\widehat{A}_t(x_i, y_i) := r(x_i, y_i) - r(x_i, y_i')$ be the estimate of $A^{\pi_{\theta_t}}(x_i, y_i)$
   - Call a `PrivUpdate` oracle on $D_t := \{(x_i, y_i, y_i', \widehat{A}_t(x_i, y_i))\}_{i=1}^m$ to find next policy θ_{t+1}
3. **End For**

Proposition: Suppose `PrivUpdate` satisfies (ε, δ)-DP under Definition of DP in PO, then Algorithm 1 satisfies (ε, δ)-DP in terms of Definition of standard DP.

Differentially Private Policy Gradient

Algorithm 2: `PrivUpdate` Instantiation for `DP-PG`

1. Compute the empirical policy gradient:
$$\widehat{\nabla}_m J(\theta) := \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \log \pi_{\theta_t}(y_i \mid x_i) \cdot \widehat{A}_t(x_i, y_i)$$
2. Add Gaussian noise: $\widetilde{g}_t := \widehat{\nabla}_m J(\theta) + \mathcal{N}(0, \sigma^2 I)$
3. Output policy: $\theta_{t+1} = \theta_t + \eta \cdot \widetilde{g}_t$

Assumption 1: (Fisher-non-degenerate, adapted from Assumption 2.1 of Ding et.al [3]) For all $\theta \in \mathbb{R}^d$, there exists $\gamma > 0$ s.t. the Fisher information matrix $F_\rho(\theta)$ induced by policy π_θ and initial state distribution ρ satisfies

$$F_\rho(\theta) = \mathbb{E}_{x\sim\rho, y\sim\pi_\theta(\cdot|x)} \left[\nabla_\theta \log \pi_\theta(y|x) \nabla_\theta \log \pi_\theta(y|x)^\top\right] \geqslant \gamma \mathbf{I}_d.$$

Assumption 2: (Compatible, adapted from Assumption 4.6 in Ding et.al [3]) For all $\theta \in \mathbb{R}^d$, there exists $\alpha_{\text{bias}} > 0$ such that the *transferred compatible function approximation error* satisfies

$$\mathbb{E}_{x\sim\rho, y\sim\pi_{\theta^*}(\cdot|s)} \left[(A^{\pi_\theta}(x, y) - u^{*\top} \nabla_\theta \log \pi_\theta(y|x))^2\right] \leqslant \alpha_{\text{bias}},$$

where π_{θ^*} is an optimal policy and $u^* = F_\rho(\theta)^\dagger \nabla J(\theta)$.

Theorem 1: For any $\alpha > 0$, `DP-PG` enjoys the following average regret guarantee

$$J^* - \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[J(\theta_t)\right] \leqslant O(\alpha) + O\left(\sqrt{\alpha_{\text{bias}}}\right),$$

when the sample size satisfies $N \geqslant O_\delta\left(\frac{1}{\alpha^4\gamma^4} + \frac{\sqrt{d}}{\alpha^3\gamma^3\varepsilon}\right)$.

Differentially Private NPG

Algorithm 3: `PrivUpdate` Instantiation for `DP-NPG`

1. Call the `PrivLS` oracle on $D_t := \{(x_i, y_i, \widehat{A}_t(x_i, y_i))\}$ to find an approximate minimizer w_t of
$$\underset{w\in\mathcal{W}}{\operatorname{argmin}}\, F_t(w) := \mathbb{E}_{x\sim\rho, y\sim\mu(\cdot|x)} \left[\left(A^{\pi_{\theta_t}}(x, y) - w^\top \nabla \log \pi_{\theta_t}(y|x)\right)^2\right]$$
2. Output policy $\theta_{t+1} = \theta_t + \eta w_t$

Assumption 3: For each $t \in [T]$, the `PrivLS` oracle satisfies (ε, δ)-DP while ensuring that with probability at least $1 - \zeta$,

$$\mathbb{E}_{x\sim\rho, y\sim\mu(\cdot|x)} \left[\left(A^{\pi_{\theta_t}}(x, y) - w_t^\top \nabla \log \pi_{\theta_t}(y|x)\right)^2\right] \leqslant \text{err}_t^2(m, \varepsilon, \delta, \zeta),$$

for some error function $\text{err}_t^2(m, \varepsilon, \delta, \zeta)$ over batch size m, privacy parameters ε, δ, and probability ζ.

Theorem 2: `DP-NPG` satisfies (ε, δ)-DP as in Definition 1. Moreover, if $\pi_1 := \pi_{\theta_1}$ is a uniform distribution at each state and $\eta = \sqrt{\frac{2\log|\mathcal{Y}|}{T\beta W^2}}$, with probability at least $1-\zeta$, for any comparator policy π^*, we have

$$J(\pi^*) - \frac{1}{T} \sum_{t=1}^{T} J(\pi_t) \leqslant \sqrt{\frac{\beta W^2 \log|\mathcal{Y}|}{2T}} + \frac{\sqrt{C_{\mu\to\pi^*}}}{T} \sum_{t=1}^{T} \text{err}_t(m, \varepsilon, \delta, \zeta),$$

where $C_{\mu\to\pi^*} := \max_{x,y} \frac{\pi^*(y|x)}{\mu(y|x)}$ and $\pi_t := \pi_{\theta_t}$.

Applications of `DP-NPG`

Exponential Mechanism

Algorithm 5: `PrivLS` Instantiation for `DP-NPG` via Exponential Mechanism

`// Input: privacy budget` ε`, current policy` θ_t`, reward range` $R_{\max}$

1. Sample $w_t \in \mathcal{W}$ with the following distribution:
$$P(w) \propto \exp\left(-\frac{\varepsilon}{8R_{\max}^2} \cdot L(w)\right) \ \forall w \in \mathcal{W},$$
where $L(w) := \sum_{i\in[m]}[w^\top \nabla \log \pi_{\theta_t}(y_i|x_i) - \widehat{A}_t(x_i, y_i)]^2$

Assumption 4: Assume the advantage function satisfies approximate realizability:

$$\inf_{w\in\mathcal{W}} \mathbb{E}_{x\sim\rho, y\sim\mu(\cdot|x)} \left[(A^{\pi_{\theta_t}}(x, y) - w^\top \nabla \log \pi_{\theta_t}(y|x))^2\right] \leqslant \alpha_{\text{approx}}. \quad (1)$$

Then, sampling $\widehat{w}$ via the exponential mechanism yields:

$$\mathbb{E}_{(x,y)\sim\rho\times\mu(\cdot|x)} \left[(\widehat{w}^\top \nabla \log \pi_{\theta_t}(y|x) - A^{\pi_{\theta_t}}(x, y))^2\right] \lesssim \frac{R^2 \log(|\mathcal{W}|/\zeta)}{m} + \frac{R^2 \log(|\mathcal{W}|/\zeta)}{\varepsilon m} + \alpha_{\text{approx}}.$$

Corollary 1: Consider `DP-NPG` with ***PrivLS*** as in Algorithm above. Then, `DP-NPG` satisfies $(\varepsilon, 0)$-DP. Suppose for each $t \in [T]$, there exists an α_{approx} such that (1) holds. Then, under the same assumptions in Theorem 2, we have

$$J(\pi^*) - \frac{1}{T} \sum_{t=1}^{T} J(\pi_t) \lesssim \sqrt{\frac{\beta W^2 \log|\mathcal{Y}|}{T}} + \sqrt{C_{\mu\to\pi^*} \alpha_{\text{approx}}} + \sqrt{C_{\mu\to\pi^*} \cdot \frac{(1 + 1/\varepsilon) \log(|\mathcal{W}|/\zeta)}{m}}.$$

This implies that, for a given suboptimality gap of $O(\alpha + \sqrt{C_{\mu\to\pi^*}\alpha_{\text{approx}}})$, the sample complexity bound is $N = T \cdot m = \widetilde{O}\left((\frac{1}{\alpha^4} + \frac{1}{\alpha^4\varepsilon}) \cdot \log|\mathcal{W}| \cdot \beta W^2\right)$.

Log-linear policy class with realizability

Corollary 2: Consider `DP-NPG` with the above log-linear class (with smoothness parameter $\beta = B^2$). Suppose ***PrivLS*** is instantiated with the ***ISSP*** algorithm in [1]. Then, by [1, Theorem 5], we have that $\text{err}_t(m, \varepsilon, \delta, \zeta) \leqslant \alpha$, when $m \geqslant \widetilde{O}\left(\frac{d}{\alpha^2} + \frac{d\sqrt{\log(1/\delta)}}{\alpha\varepsilon} + \frac{d(\log(1/\delta))^2}{\varepsilon^2}\right)$. Thus, by Theorem 2, for a suboptimality gap of $O(\alpha)$, the sample complexity bound is $N = T \cdot m = \widetilde{O}_\delta\left((\frac{d}{\alpha^4} + \frac{d}{\alpha^3\varepsilon} + \frac{d}{\alpha^2\varepsilon^2}) \cdot B^2W^2\right)$.

Corollary 3: Consider `DP-NPG` with the above log-linear class (with smoothness parameter $\beta = B^2$). Suppose ***PrivLS*** is instantiated with Algorithm 5 in [2]. Then, by [2, Theorem 6.2], we have that $\text{err}_t(m, \varepsilon, \delta, \zeta) \leqslant \alpha$ when $m \geqslant \widetilde{O}\left(\frac{\log(1/\zeta)}{\alpha^4} + \frac{\sqrt{\log(1/\zeta)\log(1/\delta)}}{\alpha^3\varepsilon}\right)$. Thus, by Theorem 2, for a suboptimality gap of $O(\alpha)$, the sample complexity bound is $N = T \cdot m = \widetilde{O}_\delta\left((\frac{1}{\alpha^6} + \frac{1}{\alpha^5\varepsilon}) \cdot B^2W^2\right)$.

References

[1] Gavin Brown, Jonathan Hayase, Samuel Hopkins, Weihao Kong, Xiyang Liu, Sewoong Oh, Juan C. Perdomo, and Adam Smith.
Insufficient statistics perturbation: Stable estimators for private least squares, 2024.

[2] Fan Chen, Jiachun Li, Alexander Rakhlin, and David Simchi-Levi.
Near-optimal private learning in linear contextual bandits, 2025.

[3] Yuhao Ding, Junzi Zhang, and Javad Lavaei.
On the global optimum convergence of momentum-based policy gradient.
In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 1910–1934. PMLR, 28–30 Mar 2022.

[4] Rui Yuan, Robert M Gower, and Alessandro Lazaric.
A general sample complexity analysis of vanilla policy gradient.
In *International Conference on Artificial Intelligence and Statistics*, pages 3332–3380. PMLR, 2022.