On the Sample Complexity of Differentially Private Policy Optimization

Yi He, Wayne State University

CAD Seminar @ WSU

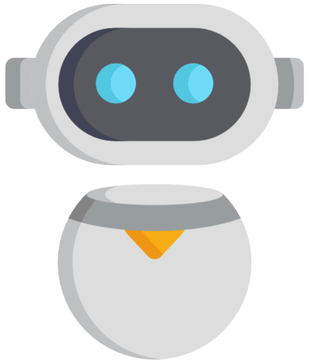
Oct 22, 2025

Joint work with Xingyu Zhou

accepted by NeurIPS 2025

Motivation

Policy Optimization - A Cornerstone of Modern RL

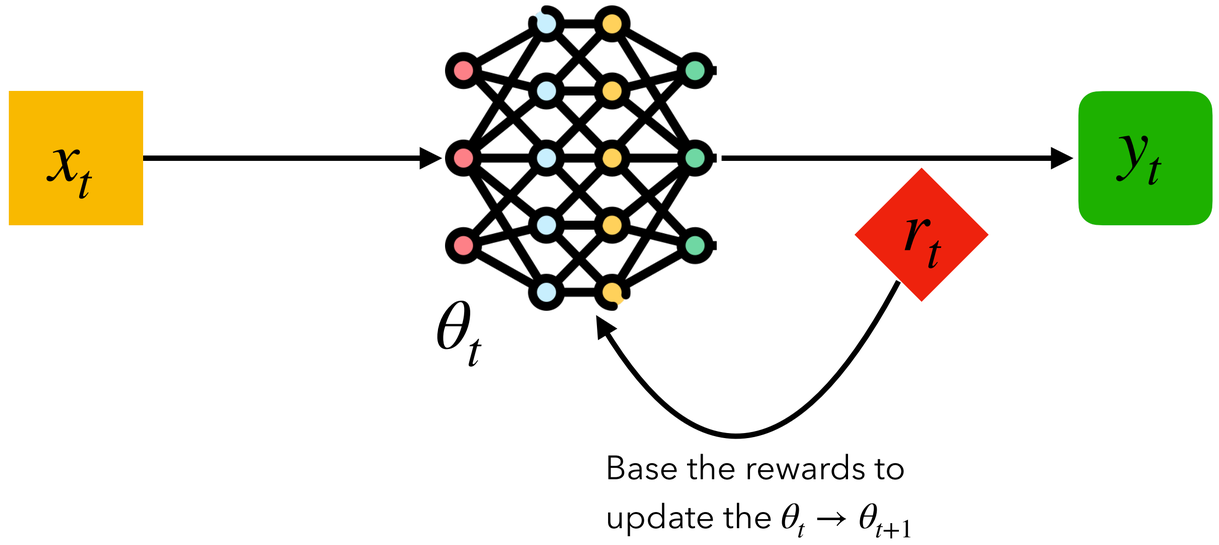


Target:

Maximize the following objective

$$J(\pi_\theta) = J(\theta) := \mathbb{E}_{x\sim\rho, y\sim\pi_\theta(\cdot|x)}\left[r(x,y)\right],$$

where ρ is some distribution of the initial state.

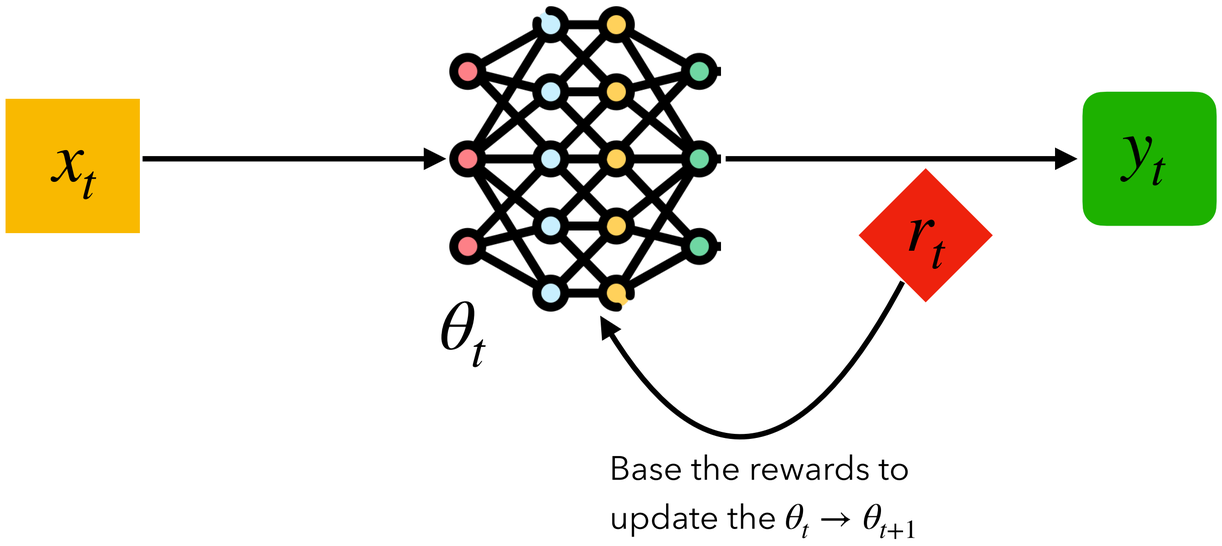


Policy Optimization - A Cornerstone of Modern RL

Target:

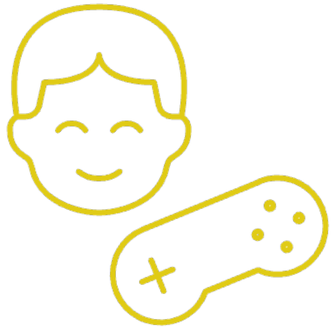
Maximize the following objective

$$J(\pi_\theta) = J(\theta) := \mathbb{E}_{x\sim\rho, y\sim\pi_\theta(\cdot|s)}\left[r(x,y)\right].$$



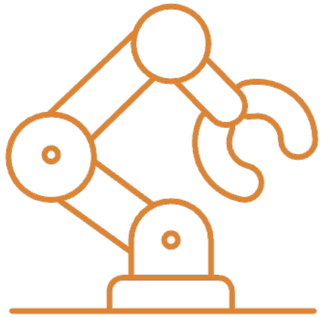
- One of the most widely used RL methods
- Directly optimizes the policy (unlike value-based methods)
- Core algorithms: PG, NPG, TRPO, PPO, GRPO
- Rich theoretical foundation established over decades

Real-World Impact Across Domains



Games

AlphaGo and other advanced game-playing AI.



Robotics

Autonomous control systems for robots.

Healthcare

AI for personalized treatment optimization.

LLMs

Training large language models like ChatGPT.

Privacy Risks in Learning from Sensitive Data

Both RL and LLMs learn from personal information – and risk leaking it.

🩺 Healthcare RL

- *State*: Patient medical history
- *Action*: Treatment decision
- *Reward*: Health outcome
- ⚠️ *Risk*: Leakage of private medical data

🧠 LLM Training

- *Input*: User prompts containing private info
- *Process*: Model training or fine-tuning
- ⚠️ *Risk*: Memorization and regurgitation of sensitive content

Both involve optimizing over sensitive data without formal privacy guarantees.

Differential Privacy - The Golden Standard

Differential Privacy (DP) has become the cornerstone of modern data protection. Proposed by *Dwork et al.*[1], it formalizes privacy guarantees by ensuring that the outcome of an algorithm is nearly unaffected by the presence or absence of any individual user.

Real-World Successes

- ✅ **Supervised Learning** — DP-SGD for private model training
- ✅ **Federated Learning** — privacy at the edge
- ✅ **Data Analytics** — deployed by Google, Apple, Microsoft
- ❓ **Reinforcement Learning** — an emerging frontier

The Research Gap

Lack of Theoretical Understanding for Privacy in Policy Optimization

Current State:

✅ Extensive PO theory for convergence and sample efficiency

✅ Broad deployment of PO in safety-critical domains (e.g., healthcare, LLM alignment)

❌ No theoretical results on *sample complexity** under differential privacy

❌ Standard DP notion fails under on-policy data generation

❌ No unified framework connecting privacy, sample complexity, and PO theory

"Without a solid theory, privacy-preserving PO remains empirical – limiting safe and reliable deployment."

* The sample complexity typically refers to the total number of sampled trajectories for finding an α-optimal policy (i.e., $J(\pi^*) - J(\hat{\pi}) \leq \alpha$).

Three Fundamental Questions for Private Policy Optimization

- 🧩 **Definition:**
  What is the right notion of differential privacy for policy optimization?

- ⚙️ **Framework:**
  How to design a **unified*** framework for private policy optimization?

- 📊 **Theory:**
  What's the sample complexity cost induced by differential privacy in PO?

* unified means using a unified algorithm to apply different algorithms.

Q1:

What is the right notion of differential privacy for policy optimization?

Differentially Privacy

Mathematical Definition

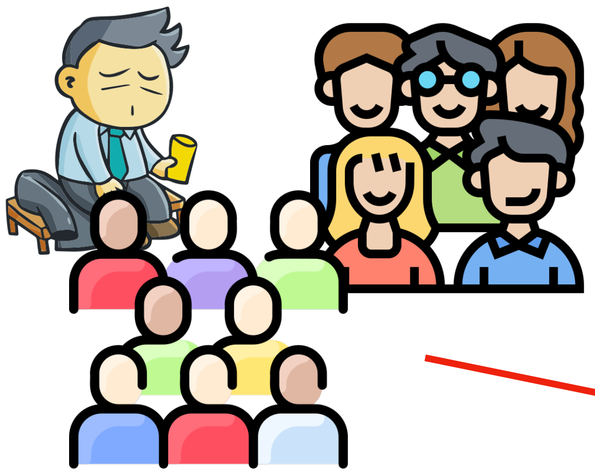
Definition 1 (Dwork et al.*[1]*)

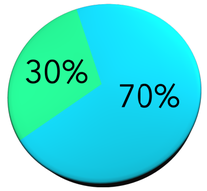
A randomized algorithm $\mathscr{M}$ is said to satisfy (ϵ, δ)-**Differential Privacy** if, for any two datasets D and D' that differ by only ***one record***, and for any possible output S of the algorithm:

$$\Pr[\mathscr{M}(D) \in S] \leq e^{\epsilon} \cdot \Pr[\mathscr{M}(D') \in S] + \delta$$

An example of healthcare(without DP)

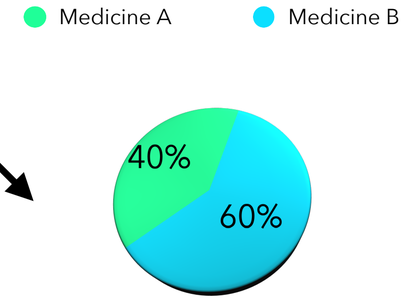
John





John in the Training Data

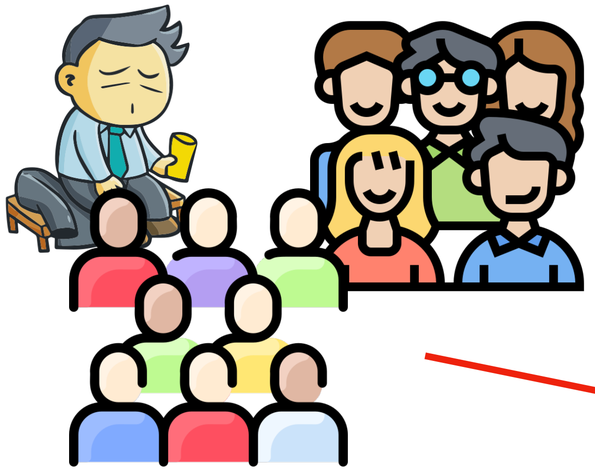
Ethan is an adversary who knows the system's outputs and almost everything except ***John,*** and also holds aggregate statistics.

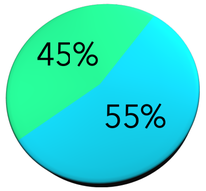


John not in the Training Data

An example of healthcare(with DP)

John



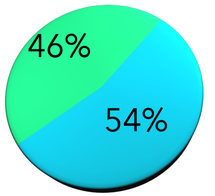


Medicine A Medicine B

John in the Training Data

Ethan is an adversary who knows the system's outputs and almost everything except ***John,*** and also holds aggregate statistics.

Medicine A Medicine B



John not in the Training Data

Differentially Privacy

Mathematical Definition

Definition 1 (Dwork et al.*[1]*)

A randomized algorithm $\mathscr{M}$ is said to satisfy (ϵ, δ)-**Differential Privacy** if, for any two datasets D and D' that differ by only ***one record***, and for any possible output S of the algorithm:

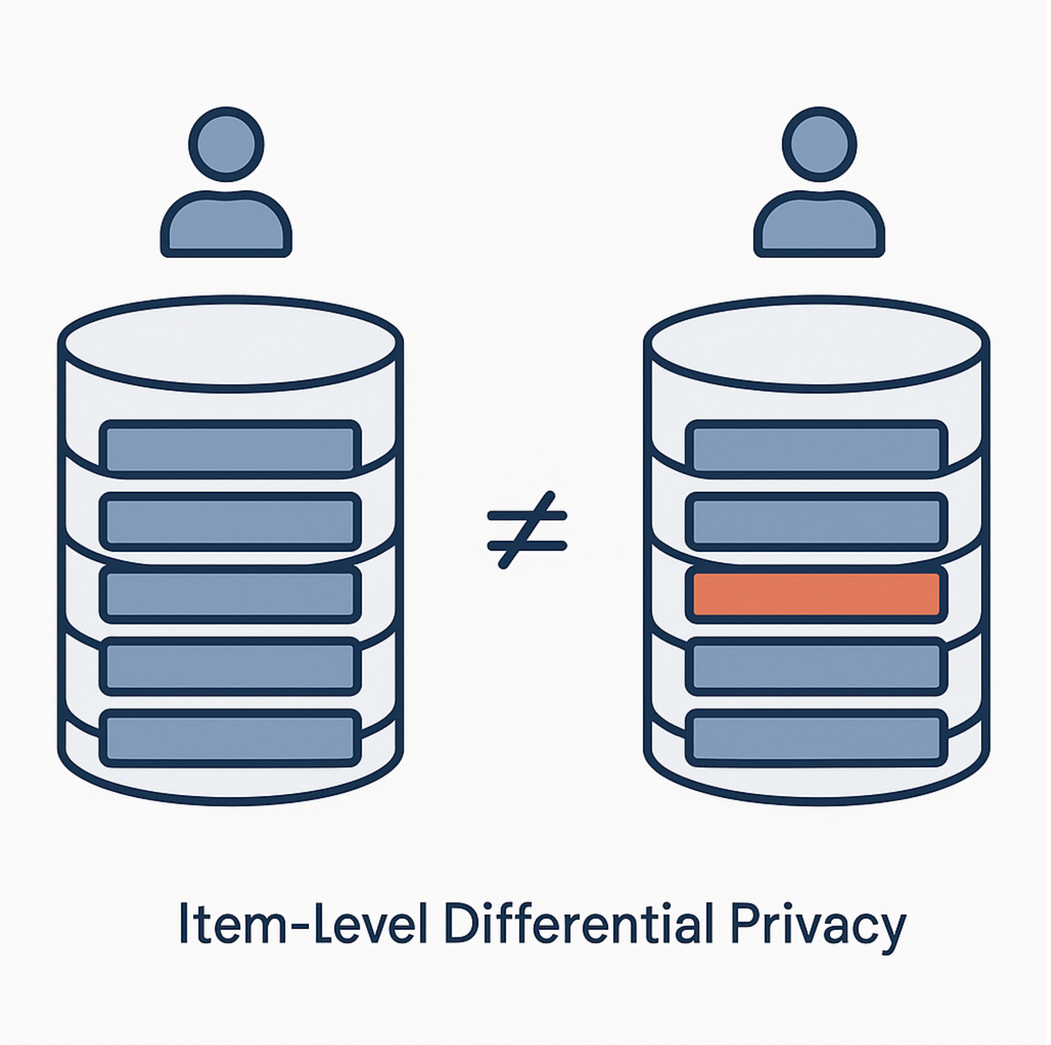
$$\Pr[\mathscr{M}(D) \in S] \leq e^{\epsilon} \cdot \Pr[\mathscr{M}(D') \in S] + \delta$$

What constitutes "one record"?

Granularities of Differentially Privacy

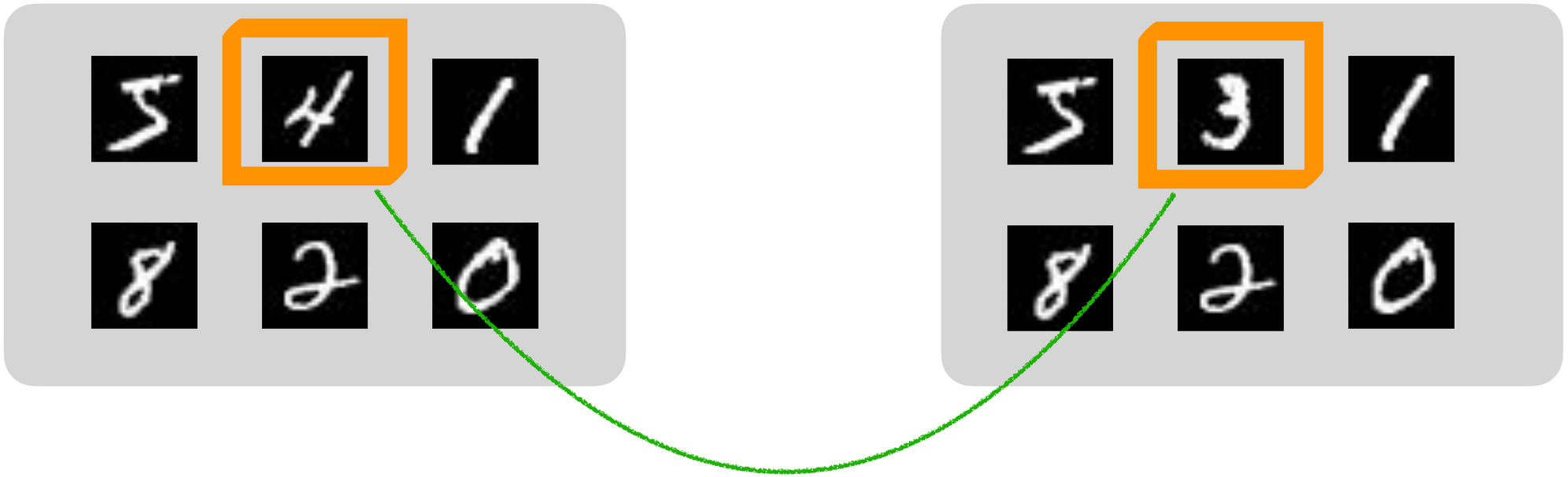
Item-Level DP

The classical setting: two datasets are neighbors if they differ in **exactly one interaction record**.



Item-Level DP

In **Supervised Learning** problem, the standard DP notion can be directly used.



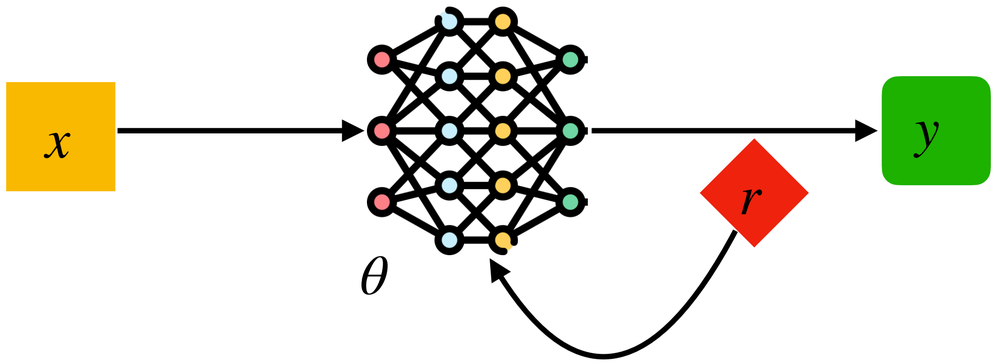
Only *one record* different.

Item-Level DP

Why not work in **Policy Optimization**?

- No such a **fixed** dataset in PO

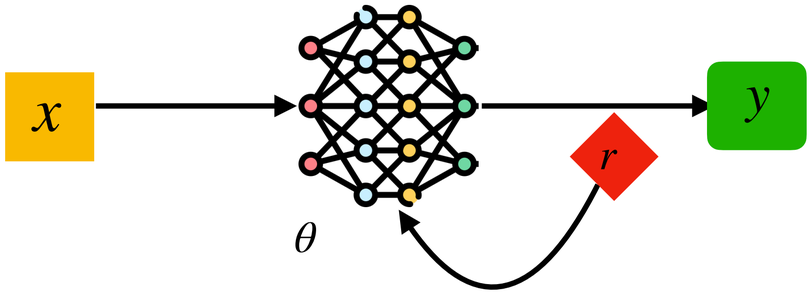
the actions are often sampled in the on-policy fashion, i.e., using the most recent policy;



PO fashion

Feature(x)	Label(y)
	5
	4
	0

Most common dataset



Item-Level DP

Why not work in **Policy Optimization**?

- No such a **fixed** dataset in PO

  the actions are often sampled in the on-policy fashion, i.e., using the most recent policy;

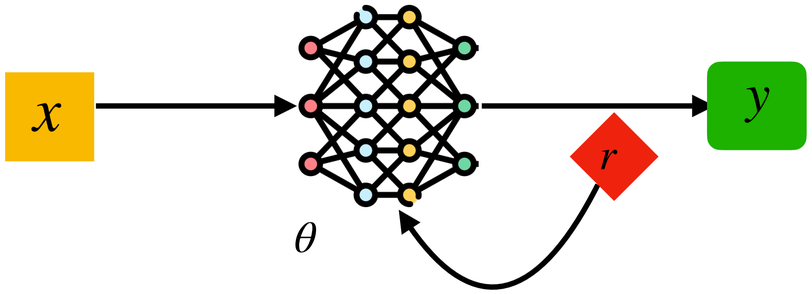
- The neighboring relation of differing in one sample (x_i, y_i) actually does not hold

  changing one sample will lead to difference in all future samples due to different policies onward.

Can this be true?

$(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4) \ldots$

$(x_1, y_1), (x_2', y_2'), (x_3, y_3), (x_4, y_4) \ldots$



Item-Level DP

Why not work in **Policy Optimization**?

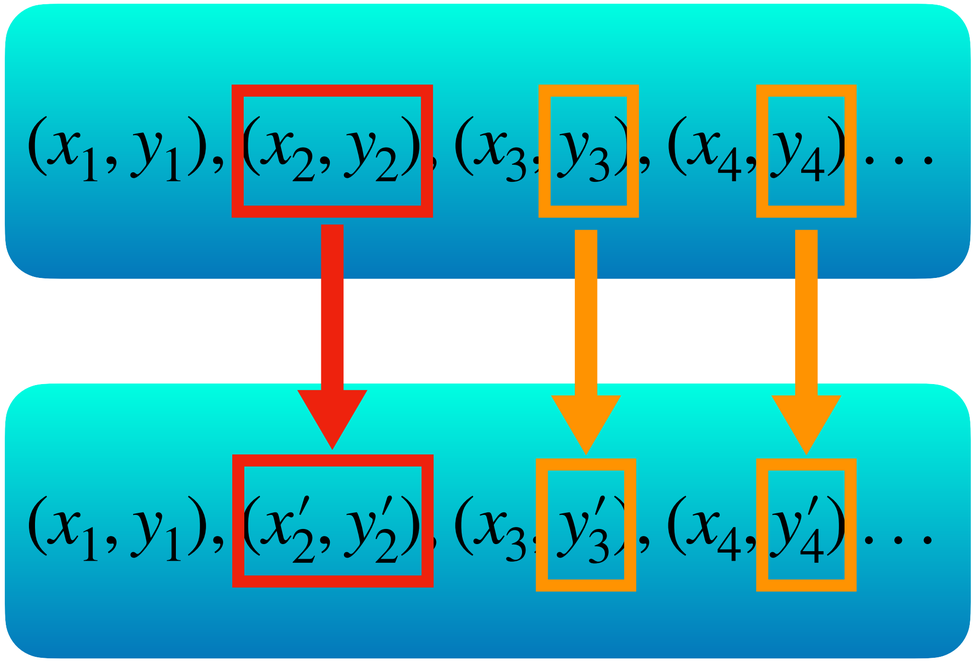
- No such a **fixed** dataset in PO

  the actions are often sampled in the on-policy fashion, i.e., using the most recent policy;

- The neighboring relation of differing in one sample (x_i, y_i) actually does not hold

  changing one sample will lead to difference in all future samples due to different policies onward.

Absolutely Not!!!



Granularities of Differentially Privacy

Definition 2 (DP in PO)

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

$$\mathbb{P}[\mathscr{M}(D) \in S] \leq e^{\epsilon} \cdot \mathbb{P}[\mathscr{M}(D') \in S] + \delta\,.$$

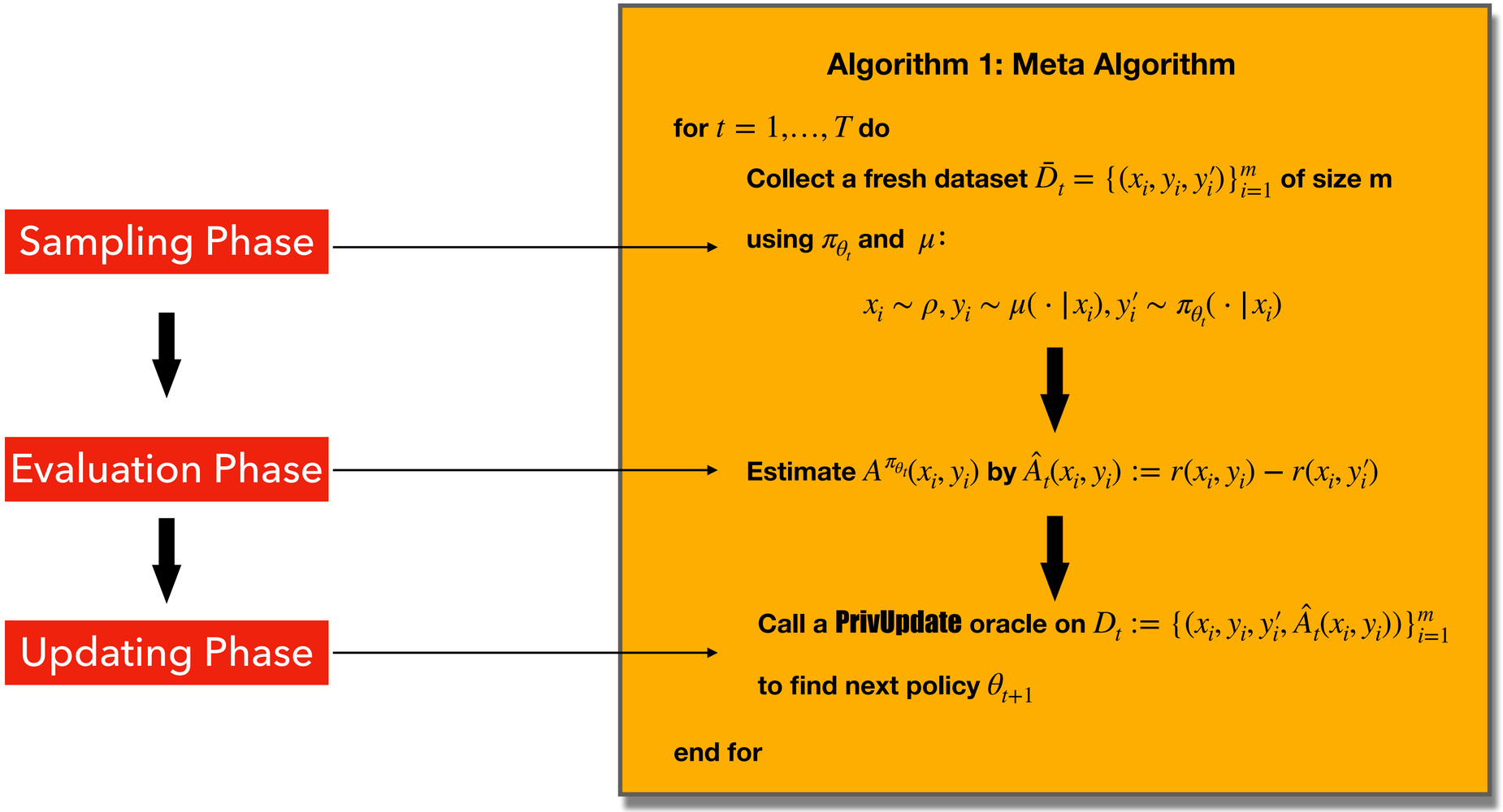
Each user can be each prompt (which is the x and it is static), but the response y and reward $r(x, y)$ are dynamic, since they are determined in the on-policy manner. This is in sharp contrast to supervised learning where the whole dataset is static and fixed in advance.

Q2:

How to design a unified framework for private policy optimization?

A Meta Algorithm for Private PO

Objective: Maximize expected reward through iterative policy parameter optimization



A Meta Algorithm for Private PO

Objective: Maximize expected reward through iterative policy parameter optimization

Algorithm 1: Meta Algorithm

for $t = 1, \ldots, T$ **do**

Collect a fresh dataset $\bar{D}_t = \{(x_i, y_i, y_i')\}_{i=1}^m$ **of size m**

Proposition 1

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

Updating Phase

Call a PrivUpdate oracle on $D_t := \{(x_i, y_i, y_i', \hat{A}_t(x_i, y_i))\}_{i=1}^m$

to find next policy θ_{t+1}

end for

Proposition 1

Suppose ***PrivUpdate*** satisfies (ϵ, δ)-DP under the standard DP definition, then Algorithm 1 also satisfies (ϵ, δ)-DP in terms of the Definition of **DP in PO**.

Proof Sketch.

This result simply follows from our **one-pass algorithm** and (adaptive) **parallel composition** of DP, by noting that changing one **user** would only change one record in D_t of a single $t \in [T]$.

Extensions of Meta Algorithm.

- The same argument can extend to **online settings**, where a stream of N "users" arrive sequentially and updates are applied adaptively.
- Our framework can also accommodate **Joint Differential Privacy (JDP)*** by the so-called *billboard lemma*[3].

* JDP guarantees that changing one "user" (say u) will not change all the actions prescribed to all other "users" except u, as well as the final policy.

Differentially Private Policy Gradient

Vanilla policy gradient (PG)[4].

(one simple and direct approach)

$$\theta_{t+1} = \theta_t + \eta \nabla J(\theta_t)$$

where $\eta > 0$ is some learning rate, $\nabla J(\theta_t)$ is the gradient at step t, and θ_1 is some initial value.

The gradient can be written as follows by the classic policy gradient theorem:

$$\nabla J(\theta) = \mathbb{E}_{x\sim\rho, y\sim\pi_\theta(\cdot|x)} \left[A^{\pi_\theta}(x, y) \nabla_\theta \log \pi_\theta(y \,|\, x) \right].$$

Algorithm 2: Differentially Private Policy Gradient

for $t = 1, \ldots, T$ **do**

Collect a fresh dataset $\bar{D}_t = \{(x_i, y_i, y_i')\}_{i=1}^m$ **of size m**

using π_{θ_t} **and** μ:

$$x_i \sim \rho, y_i \sim \mu(\cdot \,|\, x_i), y_i' \sim \pi_{\theta_t}(\cdot \,|\, x_i)$$

Estimate $A^{\pi_{\theta_t}}(x_i, y_i)$ **by** $\hat{A}_t(x_i, y_i) := r(x_i, y_i) - r(x_i, y_i')$

Compute gradient:

$$\hat{\nabla}_m J(\theta) := \frac{1}{m} \sum_{i=1}^m \nabla_\theta \log \pi_{\theta_t}(y_i \,|\, x_i) \cdot \hat{A}_t(x_i, y_i)$$

Add noise:

$$\widetilde{g}_t := \hat{\nabla}_m J(\theta) + \mathcal{N}(0, \sigma^2 I)$$

Output policy: $\theta_{t+1} = \theta_t + \eta \cdot \widetilde{g}_t$

end for

Differentially Private NPG

Natural policy gradient (NPG)[5].

(uses the Fisher information matrix)

$$\theta_{t+1} = \theta_t + \eta F_\rho^\dagger(\theta_t)\nabla J(\theta_t)$$

An equivalent way to write the above update is

$$\theta_{t+1} = \theta_t + \eta \cdot w_t,$$

$$w_t \in \arg\min_w \mathbb{E}_{x\sim\rho, y\sim\pi_{\theta_t}(\cdot|x)}\left[\left(A^{\pi_{\theta_t}}(x,y) - w^\top \nabla \log \pi_{\theta_t}(y\,|\,x)\right)^2\right],$$

which essentially reduces PO to a sequence of regression problems.

Algorithm 3: Differentially Private NPG

for $t = 1, \ldots, T$ **do**

Collect a fresh dataset $\bar{D}_t = \{(x_i, y_i, y_i')\}_{i=1}^m$ **of size m**

using π_{θ_t} **and** μ**:**

$$x_i \sim \rho, y_i \sim \mu(\cdot\,|\,x_i), y_i' \sim \pi_{\theta_t}(\cdot\,|\,x_i)$$

Estimate $A^{\pi_{\theta_t}}(x_i, y_i)$ **by** $\hat{A}_t(x_i, y_i) := r(x_i, y_i) - r(x_i, y_i')$

- Call the **PrivLS** oracle on $D_t := \{(x_i, y_i, \hat{A}_t(x_i, y_i))\}$ to find an approximate minimizer w_t of

$$w_t = \arg\min_w \mathbb{E}\left[\left(A^{\pi_{\theta_t}}(x,y) - w^\top \nabla \log \pi_{\theta_t}(y\,|\,x)\right)^2\right]$$

- Output policy: $\theta_{t+1} = \theta_t + \eta w_t$

end for

Differentially Private REBEL

Regression to Relative Reward Based RL (REBEL)[6].

$$\theta_{t+1} = \arg\min_{\theta} \mathbb{E}\left[\frac{1}{\eta}\left(\ln\frac{\pi_\theta(y \mid x)}{\pi_{\theta_t}(y \mid x)} - \ln\frac{\pi_\theta(y' \mid x)}{\pi_{\theta_t}(y' \mid x)}\right) - \big(r(x,y) - r(x,y')\big)\right]^2$$

where the expectation here is over $x \sim \rho, y \sim \mu(\cdot\,|x), y' \sim \pi_{\theta_t}(\cdot\,|x)$, and μ can be either on-policy distribution π_{θ_t} or any offline reference policy.

Algorithm 4: Differentially Private REBEL

for $t = 1, \ldots, T$ **do**

Collect a fresh dataset $\bar{D}_t = \{(x_i, y_i, y_i')\}_{i=1}^m$ **of size m**

using π_{θ_t} **and** μ**:**

$$x_i \sim \rho, y_i \sim \mu(\cdot\,|x_i), y_i' \sim \pi_{\theta_t}(\cdot\,|x_i)$$

Estimate $A^{\pi_{\theta_t}}(x_i, y_i)$ **by** $\hat{A}_t(x_i, y_i) := r(x_i, y_i) - r(x_i, y_i')$

- Call the **PrivLS** oracle on $D_t := \{(x_i, y_i, \hat{A}_t(x_i, y_i))\}$ to find an approximate minimizer w_t of

$$\arg\min_{\theta \in \Theta} F_t(\theta) = \mathbb{E}\left[\frac{1}{\eta}\left(\ln\frac{\pi_\theta(y\,|x)}{\pi_{\theta_t}(y\,|x)} - \ln\frac{\pi_\theta(y'\,|x)}{\pi_{\theta_t}(y'\,|x)}\right) - \left(\hat{A}_t(x_i, y_i)\right)\right]^2,$$

- Output policy: $\theta_{t+1} = \theta_t + \eta w_t$

end for

Q3:

What's the sample complexity cost induced by differential privacy in PO?

Differentially Private Policy Gradient

Theorem 1 **Privacy guarantee**

Assume for any $x \in \mathcal{X}$ and $\theta \in \Theta$, there exists a constant G such that $\|\nabla_\theta \log \pi_\theta(y \mid x)\| \leq G$. Then, setting $\sigma^2 = \frac{16 \log(1.25/\delta) R_{\max}^2 G^2}{m^2 \epsilon^2}$ in Algorithm 2 (**DP-PG**) ensures that **DP-PG** satisfies (ϵ, δ)-DP, as in DP in PO.

Compute gradient:

$$\hat{\nabla}_m J(\theta) := \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \log \pi_{\theta_t}(y_i \mid x_i) \cdot \hat{A}_t(x_i, y_i)$$

Add noise:

$$\tilde{g}_t := \hat{\nabla}_m J(\theta) + \mathcal{N}(0, \sigma^2 I)$$

Output policy: $\theta_{t+1} = \theta_t + \eta \cdot \tilde{g}_t$

Assumption 1* **Lipschitz Smoothness (LS)**[7]

There exist constants $G, F > 0$ such that for every state $x \in \mathcal{X}$, the gradient and Hessian of $\log \pi_\theta(\cdot \mid x)$ of any $\theta \in \Theta$ satisfy

$\|\nabla_\theta \log \pi_\theta(y \mid x)\| \leq G$ and $\|\nabla_\theta^2 \log \pi_\theta(y \mid x)\| \leq F$.

* Both tabular and log-linear softmax satisfy this assumption.

Theorem 2 **First-Order Stationary Point Convergence**

Under Assumption 1, there exists a proper parameter choices of m and η , such that **DP-PG** achieves

$$\mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] \leq O_\delta\left(\frac{1}{\sqrt{N}} + \left(\frac{\sqrt{d}}{N\epsilon}\right)^{2/3}\right),$$

where θ_U is uniformly sampled from $\{\theta_1, \ldots, \theta_T\}$, and $N = T \cdot m$.

Differentially Private Policy Gradient

Theorem 2 **First-Order Stationary Point Convergence**

Under Assumption 1, there exists a proper parameter choices of m and η , such that **DP-PG** achieves

$$\mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] \leq O_\delta\left(\frac{1}{\sqrt{N}} + \left(\frac{\sqrt{d}}{N\epsilon}\right)^{2/3}\right),$$

where θ_U is uniformly sampled from $\{\theta_1, \ldots, \theta_T\}$, and $N = T \cdot m$.

For a fixed N, the key here is to balance between batch size m and number of iterations T so as to balance between the **per-iteration accuracy** and the **total number of updates**. This balance, in turn, depends on the specific choice of **PrivUpdate** oracle, which will be instantiated in the next sections for **DP-PG**, **DP-NPG**, and **DP-REBEL**, respectively.

Differentially Private Policy Gradient

We now turn our focus to the global optimum convergence in the sense of average regret

Assumption 2 **Fisher-non-degenerate**[8]

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

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

This assumption is **standard** in the literature on non-private PG methods and holds for many common policy classes, such as **Gaussian policies** and even **certain neural network** policies

The function $\nabla_\theta \log \pi_\theta(y|x)$ is "**isotropically balanced and sufficiently informative**" over the (x, y) distribution induced by ρ and the current policy. This prevents situations where there is a complete lack of signal in certain directions.

Assumption 3 **Compatible**[8]

For all $\theta \in \mathbb{R}^d$, there exists $\alpha_{\text{bias}} > 0$ such that the transferred compatible function approximation error satisfies

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

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

When using the "compatible" features $\nabla_\theta \log \pi_\theta(y|x)$ to approximate the advantage function, the error, when transferred to the distribution of the optimal policy, is upper-bounded.

This is also a common assumption in the PG literature to handle function approximation error in the non-tabular case.

Differentially Private Policy Gradient

Assumption 1* **Lipschitz Smoothness (LS)**[7]

There exist constants $G, F > 0$ such that for every state $x \in \mathcal{X}$, the gradient and Hessian of $\log \pi_\theta(\cdot \mid x)$ of any $\theta \in \Theta$ satisfy

$$\|\nabla_\theta \log \pi_\theta(y \mid x)\| \le G \text{ and } \|\nabla^2_\theta \log \pi_\theta(y \mid x)\| \le F.$$

Assumption 2 **Fisher-non-degenerate**[8]

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

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

Assumption 3 **Compatible**[8]

For all $\theta \in \mathbb{R}^d$, there exists $\alpha_{\text{bias}} > 0$ such that the transferred compatible function approximation error satisfies

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

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

Lemma 1[8]

If the policy class π_θ satisfies these assumptions, then we have

$$J^* - J(\theta) \le \frac{G}{\gamma} \left\| \nabla J(\theta) \right\| + \sqrt{\alpha_{\text{bias}}}.$$

Match the FOSP!

Theorem 3

For any $\alpha > 0$, **DP-PG** enjoys the following average regret guarantee

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

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

Match the non-private case!

Recall our DP-NPG

Natural policy gradient (NPG)[5].

(uses the Fisher information matrix)

$$\theta_{t+1} = \theta_t + \eta F_\rho^\dagger(\theta_t)\nabla J(\theta_t)$$

An equivalent way to write the above update is

$$\theta_{t+1} = \theta_t + \eta \cdot w_t,$$

$$w_t \in \arg\min_w \mathbb{E}_{x\sim\rho, y\sim\pi_{\theta_t}(\cdot|x)}\left[\left(A^{\pi_{\theta_t}}(x,y) - w^\top \nabla \log \pi_{\theta_t}(y\,|\,x)\right)^2\right],$$

which essentially reduces PO to a sequence of regression problems.

Algorithm 3: Differentially Private NPG

for $t = 1, \ldots, T$ **do**

Collect a fresh dataset $\bar{D}_t = \{(x_i, y_i, y_i')\}_{i=1}^m$ **of size m**

using π_{θ_t} **and** μ**:**

$$x_i \sim \rho, y_i \sim \mu(\cdot\,|\,x_i), y_i' \sim \pi_{\theta_t}(\cdot\,|\,x_i)$$

Estimate $A^{\pi_{\theta_t}}(x_i, y_i)$ **by** $\hat{A}_t(x_i, y_i) := r(x_i, y_i) - r(x_i, y_i')$

- Call the **PrivLS** oracle on $D_t := \{(x_i, y_i, \hat{A}_t(x_i, y_i))\}$ to find an approximate minimizer w_t of

$$w_t = \arg\min_w \mathbb{E}\left[\left(A^{\pi_{\theta_t}}(x,y) - w^\top \nabla \log \pi_{\theta_t}(y\,|\,x)\right)^2\right]$$

- Output policy: $\theta_{t+1} = \theta_t + \eta w_t$

end for

Differentially Private NPG

To start with, we assume that the approximate minimizer w_t returned by **PrivLS** at each iteration satisfies:

Assumption 3 **Private estimation error**

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

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

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

Note that this assumption is **algorithm-free**, meaning one can insert any private regression implementation as long as one provides its sample error upper bound.

Assumption 4* **β-smoothness and boundedness**

$\log \pi_\theta(y\,|\,x)$is a β-smooth function of θ for all x, y, i.e.,

$$\left\| \nabla_\theta \log \pi_\theta(y\,|\,x) - \nabla_{\theta'} \log \pi_{\theta'}(y\,|\,x) \right\|_2 \leq \beta \left\| \theta - \theta' \right\|_2.$$

Moreover, there exists a constant $W > 0$ such that for all $t \in [T]$, the weight vectors w_t generated by the update rule satisfy $\|w_t\|_2 \leq W$.

Smoothness allows upper bound $\langle \theta_{t+1} - \theta_t, \nabla \log \pi_{\theta_t} \rangle$ by the log-likelihood ratio $\log \frac{\pi_{\theta_{t+1}}}{\pi_{\theta_t}}$ plus a quadratic term, thereby enabling a telescoping sum over T.

* Assumption 4 is also a standard regularity assumption commonly used even in the non-private case.

Differentially Private NPG

Assumption 3 **Private estimation error**

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

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

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

Assumption 4 **β-smoothness and boundedness**

$\log \pi_\theta(y\,|\,x)$is a β-smooth function of θ for all x, y, i.e.,

$$\left\| \nabla_\theta \log \pi_\theta(y\,|\,x) - \nabla_{\theta'} \log \pi_{\theta'}(y\,|\,x) \right\|_2 \le \beta \left\| \theta - \theta' \right\|_2.$$

Moreover, there exists a constant $W > 0$ such that for all $t \in [T]$, the weight vectors w_t generated by the update rule satisfy $\|w_t\|_2 \le W$.

Theorem 4 (Master theorem)

Let Assumption 3 and 4 hold, **DP-NPG** satisfies (ϵ, δ)-DP as in **DP in PO**. Moreover, if $\pi_1 := \pi_{\theta_1}$ is a uniform distribution at each state and $\eta = \sqrt{\dfrac{2\log|\mathscr{Y}|}{T\beta W^2}}$, with probability at least $1-\zeta$, for any comparator policy π^*, we have

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

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

* We use the most intuitive coverage definition for $C_{\mu\to\pi^*}$, i.e., the density ratio, for illustrating the key idea. One can easily extend it to other types of coverage

Differentially Private NPG

Theorem 4 (Master theorem)

Let Assumption 3 and 4 hold, **DP-NPG** satisfies (ϵ, δ)-DP as in **DP in PO**. 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

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

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

Assumption 3 **Private estimation error**

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

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

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

- Call the **PrivLS** oracle on $D_t := \{(x_i, y_i, \hat{A}_t(x_i, y_i))\}$ to find an approximate minimizer w_t of

$$w_t = \arg\min_w \mathbb{E}\left[\left(A^{\pi_{\theta_t}}(x,y) - w^\top \nabla \log \pi_{\theta_t}(y|x)\right)^2\right]$$

- Output policy: $\theta_{t+1} = \theta_t + \eta w_t$

Only need to determine the estimation error under different types of **PrivLS!**

Differentially Private NPG

Approach 1: *Exponential Mechanism*

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

- Input: $D_t = \{(x_i, y_i, \hat{A}_t(x_i, y_i))\}_{i=1}^m$, privacy budget ϵ, current policy θ_t, reward range R_{max}
- Output: w_t
- Sample $w_t \in \mathscr{W}$ with the following distribution

$$P(w) \propto \exp\left(-\frac{\epsilon}{8R_{\text{max}}^2} \cdot L(w)\right) \quad \forall w \in \mathscr{W},$$

where $L(w) := \sum_{i\in[m]} [w^\top \nabla \log \pi_{\theta_t}(y_i \,|\, x_i) - \hat{A}_t(x_i, y_i)]^2$.

Lemma 2

Assume the advantage function satisfies approximate realizability:

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

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

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

Applications of Differentially Private NPG

Approach 1: *Exponential Mechanism*

Lemma 2

Assume the advantage function satisfies approximate realizability:

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

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

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

Corollary 1 **General function class**

Consider **DP-NPG** with **PrivLS** as in Algorithm above. Then, **DP-NPG** satisfies $(\epsilon,0)$-DP. Suppose for each $t \in [T]$, there exists such an α_{approx}. Then, under the same assumptions in Theorem 4, we have

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

This implies that, for a given suboptimality gap of $O(\alpha + \sqrt{C_{\mu\to\pi^*}\alpha_{\text{approx}}})$, the sample complexity bound is

$$N = T\cdot m = \widetilde{O}\left((\frac{1}{\alpha^4} + \frac{1}{\alpha^4\epsilon})\cdot\log|\mathscr{W}|\cdot\beta W^2\right).$$

Applications of Differentially Private NPG

Approach 2: *Log-Liner Policy with realizable rewards*

When the policy is log-linear and reward r is realizable, Assumption 3 reduces to estimation error of linear regression:

$$\mathbb{E}_{x\sim\rho,y\sim\mu(\cdot|x)}\left[\left(\langle w_t - w^*, \bar{\phi}^t_{x,y}\rangle\right)^2\right] \le \mathrm{err}_t^2(m,\epsilon,\delta,\zeta)$$

where $\bar{\phi}^t_{x,y} := \phi_{x,y} - \mathbb{E}_{y'\sim\pi_{\theta_t}(\cdot|x)}[\phi_{x,y'}]$, which depends on the current policy.

Corollary 2 **Log-linear policy in low-dimension**

Consider **DP-NPG** with the above log-linear class (with smoothness parameter $\beta = B^2$). Suppose **PrivLS** is instantiated with the **ISSP** algorithm in Brown et al.[9]. Then, we have that $\mathrm{err}_t(m,\epsilon,\delta,\zeta) \le \alpha$, when

$$m \ge \widetilde{O}\left(\frac{d}{\alpha^2} + \frac{d\sqrt{\log(1/\delta)}}{\alpha\epsilon} + \frac{d(\log(1/\delta))^2}{\epsilon^2}\right).$$

Thus, by Theorem 4, for a suboptimality gap of $O(\alpha)$, the sample complexity bound is

$$N = T\cdot m = \widetilde{O}_\delta\left((\frac{d}{\alpha^4} + \frac{d}{\alpha^3\epsilon} + \frac{d}{\alpha^2\epsilon^2})\cdot B^2W^2\right).$$

One can directly replace this algorithm with more efficient methods when available.

Applications of Differentially Private NPG

Approach 2: *Log-Liner Policy with realizable rewards*

When the policy is log-linear and reward r is realizable, Assumption 3 reduces to estimation error of linear regression:

$$\mathbb{E}_{x\sim\rho,y\sim\mu(\cdot|x)}\left[\left(\langle w_t - w^*, \bar{\phi}^t_{x,y}\rangle\right)^2\right] \le \mathrm{err}_t^2(m,\epsilon,\delta,\zeta)$$

where $\bar{\phi}^t_{x,y} := \phi_{x,y} - \mathbb{E}_{y'\sim\pi_{\theta_t}(\cdot|x)}[\phi_{x,y'}]$, which depends on the current policy.

Corollary 3 **Log-linear policy in high-dimension**

Consider **DP-NPG** with the above log-linear class (with smoothness parameter $\beta = B^2$). Suppose **PrivLS** is instantiated with JDP_Improper_BatchSGD algorithm in Chen et al.[10]. Then, we have that $\mathrm{err}_t(m,\epsilon,\delta,\zeta) \le \alpha$, when

$$m \ge \widetilde{O}\left(\frac{\log(1/\zeta)}{\alpha^4} + \frac{\sqrt{\log(1/\zeta)\log(1/\delta)}}{\alpha^3\epsilon}\right).$$

Thus, by Theorem 4, for a suboptimality gap of $O(\alpha)$, the sample complexity bound is

$$N = T\cdot m = \widetilde{O}_\delta\left((\frac{1}{\alpha^6} + \frac{1}{\alpha^5\epsilon})\cdot B^2W^2\right).$$

Same as before, this algorithm can be directly replaced if more efficient methods are available.

Conclusion

Three Fundamental Questions for Private Policy Optimization

- 🧩 **Definition:**

  What is the right notion of differential privacy for policy optimization?

  We defined ***DP in PO***.

- ⚙️ **Framework:**

  How to design a unified framework for private policy optimization?

- 📊 **Theory:**

  What's the sample complexity cost induced by differential privacy in PO?

Three Fundamental Questions for Private Policy Optimization

- 🧩 **Definition:**

  What is the right notion of differential privacy for policy optimization?

- ⚙️ **Framework:**

  How to design a unified framework for private policy optimization?

  We present a **meta algorithm** for private PO, which builds upon a unified view of PG, NPG, and REBEL.

- 📊 **Theory:**

  What's the sample complexity cost induced by differential privacy in PO?

Three Fundamental Questions for Private Policy Optimization

- 🧩 **Definition:**
  What is the right notion of differential privacy for policy optimization?
- ⚙️ **Framework:**
  How to design a unified framework for private policy optimization?
- 📊 **Theory:**
  What's the sample complexity cost induced by differential privacy in PO?

  We demonstrate that privacy costs can often manifest as lower-order terms in the sample complexity.

References

[1] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3, pages 265–284. Springer, 2006.

[2] McMahan, H. B., Ramage, D., Talwar, K., and Zhang, L. (2018). *Learning Differentially Private Recurrent Language Models.* ICLR 2018.

[3] J Hsu, Z Huang, A Roth, T Roughgarden, and ZS Wu. Private matchings and allocations. SIAM Journal on Computing, 45:1953–1984, 2016.

[4] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. Advances in neural information processing systems, 12, 1999.

[5] Sham M Kakade. A natural policy gradient. Advances in neural information processing systems, 14, 2001.

[6] Zhaolin Gao, Jonathan D. Chang, Wenhao Zhan, Owen Oertell, Gokul Swamy, Kianté Brantley, Thorsten Joachims, J. Andrew Bagnell, Jason D. Lee, and Wen Sun. Rebel: Reinforcement learning via regressing relative rewards, 2024.

[7] 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.

[8] Yuhao Ding, Junzi Zhang, and Javad Lavaei. On the global optimum convergence of momentumbased 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.

[9] 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. arXiv preprint arXiv:2404.15409, 2024.

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

Q&A

Thank you!