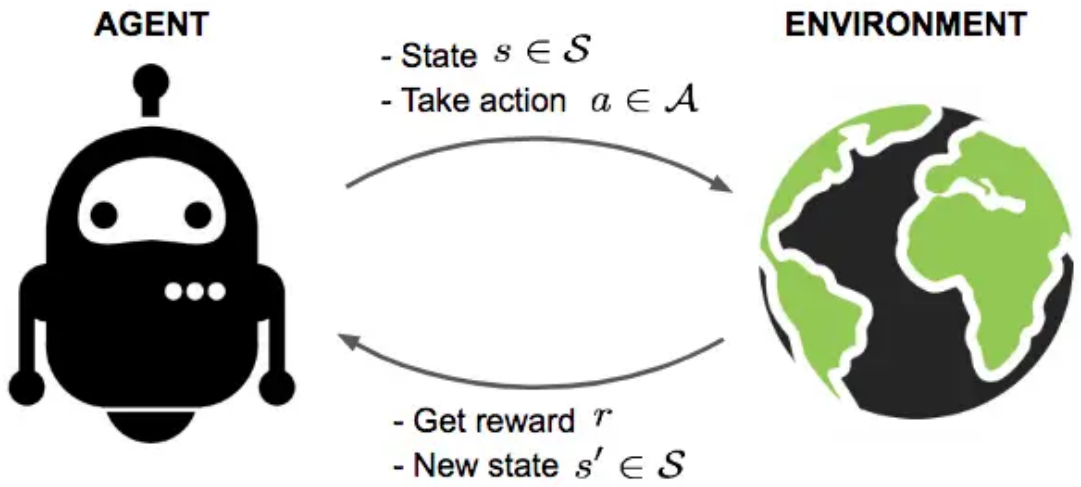
Tackling Heavy-Tailed Rewards in Reinforcement Learning with Function Approximation: Minimax Optimal and Instance-Dependent Regret Bounds

Jiayi Huang Han Zhong Liwei Wang Lin F. Yang
Peking University UCLA

NeurIPS 2023

Motivation and Background

- In real world, the states and action space can be infinite.
- We consider linear MDPs.
- Linear bandits is the simplest case of linear MDPs.



Heavy-Tailed Rewards v.s. Bounded Rewards

- Current literature assumes that rewards are either uniformly bounded or sub-Gaussian.
- Therefore, their algorithms can't handle heavy-tailed behaviors in real-world problems.
- To name a few:
  - stock returns in financial markets
  - advertiser values in online advertising

Our Contributions

Linear Bandits

Algorithm	Regret	Instance-dependent?	Minimax Optimal?	Deterministic-Optimal?	Heavy-Tailed Rewards?
OFUL (Abbasi-Yadkori et al., 2011)	$\widetilde{O}\left(d\sqrt{T}\right)$	No	**Yes**	No	No
IDS-UCB (Kirschner and Krause, 2018) Weighted OFUL+ (Zhou and Gu, 2022) AdaOFUL (Li and Sun, 2023)	$\widetilde{O}\left(d\sqrt{\sum_{t=1}^{T}\sigma_t^2}+d\right)$	**Yes**	**Yes**	**Yes**	No No $\epsilon=1$
MENU and TOFU (Shao et al., 2018)	$\widetilde{O}\left(dT^{\frac{1}{1+\epsilon}}\right)$	No	**Yes**	No	**Yes**
HEAVY-OFUL (**Ours**)	$\widetilde{O}\left(dT^{\frac{1-\epsilon}{2(1+\epsilon)}}\sqrt{\sum_{t=1}^{T}\nu_t^2}+dT^{\frac{1-\epsilon}{2(1+\epsilon)}}\right)$	**Yes**	**Yes**	**Yes**	**Yes**

Our Contributions

Linear MDPs

Algorithm	Regret	Central Moment-Dependent?	First-Order?	Minimax Optimal?	Computationally Efficient?	Heavy-Tailed Rewards?
LSVI-UCB(Jin et al., 2020)	$\widetilde{O}\left(\sqrt{d^3H^4K}\right)$	No	No	No	**Yes**	No
FORCE (Wagenmaker et al., 2022)	$\widetilde{O}\left(\sqrt{d^3H^3V_1^*K}\right)$	No	**Yes**	No	No	No
VOQL (Agarwal et al., 2023) LSVI-UCB++ (He et al., 2023)	$\widetilde{O}\left(d\sqrt{H^3K}\right)$	No	No	**Yes**	**Yes**	No
VARA (Li and Sun, 2023)	$\widetilde{O}\left(d\sqrt{H\mathcal{G}^*K}\right)$	**Yes**	**Yes**	**Yes**	**Yes**	$\epsilon = 1$
HEAVY-LSVI-UCB (**Ours**)	$\widetilde{O}\left(d\sqrt{H\mathcal{U}^*}K^{\frac{1}{1+\epsilon}} + d\sqrt{H\mathcal{V}^*K}\right)$	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Adaptive Huber Regression

- Solve θ_t as follows

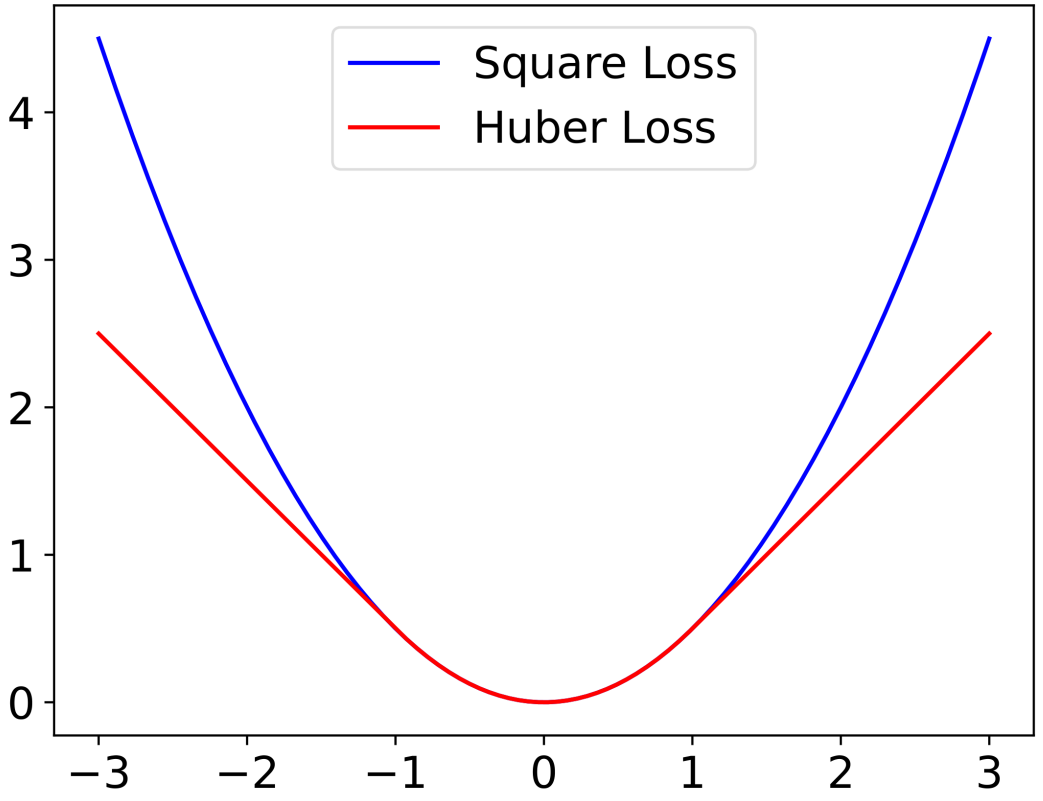
$$\theta_t = \arg\min_{\theta} L_t(\theta)$$

where $L_t(\theta) = \frac{\lambda}{2}\|\theta\|^2 + \sum_{s=1}^{t} \ell_{\tau_s}\left(\frac{R_t - \langle \phi_s, \theta \rangle}{\sigma_s}\right)$

$\ell_\tau(x)$ is Huber loss (Sun et al., 2020), which is defined as

$$\ell_\tau(x) = \begin{cases} \frac{x^2}{2} & \text{if } |x| \leqslant \tau, \\ \tau|x| - \frac{\tau^2}{2} & \text{if } |x| > \tau. \end{cases}$$

When $\tau = \infty$, it is reduced to weighted regularized least-squares regression



Linear Bandits

Algorithm 2 HEAVY-OFUL

Require: Number of total rounds T, confidence level δ, regularization parameter λ, $\sigma_{\min}$, parameters for adptive Huber regression c_0, c_1, τ_0, confidence radius β_t.

$\kappa = d\log(1 + \frac{TL^2}{d\lambda\sigma_{\min}^2})$, $\mathcal{C}_0 = \mathcal{B}_d(B)$, $\boldsymbol{H}_0 = \lambda \boldsymbol{I}$.

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

  Observe $\mathcal{D}_t$.

  Set $(\boldsymbol{\phi}_t, \cdot) = \text{argmax}_{\boldsymbol{\phi} \in \mathcal{D}_t, \boldsymbol{\theta} \in \mathcal{C}_{t-1}} \langle \boldsymbol{\phi}, \boldsymbol{\theta} \rangle$.

  Play $\boldsymbol{\phi}_t$ and observe R_t, ν_t.

  Set $\sigma_t = \max\left\{\nu_t, \sigma_{\min}, \frac{\|\boldsymbol{\phi}_t\|_{\boldsymbol{H}_{t-1}^{-1}}}{c_0}, \frac{\sqrt{LB}}{c_1^{\frac{1}{4}}(2\kappa)^{\frac{1}{4}}} \|\boldsymbol{\phi}_t\|_{\boldsymbol{H}_{t-1}^{-1}}^{\frac{1}{2}}\right\}$.

  Set $\tau_t = \tau_0 \frac{\sqrt{1+w_t^2}}{w_t} t^{\frac{1-\epsilon}{2(1+\epsilon)}}$ with $w_t = \|\boldsymbol{\phi}_t/\sigma_t\|_{\boldsymbol{H}_{t-1}^{-1}}$.

  Update $\boldsymbol{H}_t = \boldsymbol{H}_{t-1} + \sigma_t^{-2} \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\top$.

  Solve for $\boldsymbol{\theta}_t$ by Algorithm 1 and set $\mathcal{C}_t = \{\boldsymbol{\theta} \in \mathbb{R}^d | \|\boldsymbol{\theta} - \boldsymbol{\theta}_t\|_{\boldsymbol{H}_t} \leq \beta_t\}$.

end for

Result of Linear Bandits

- Regret:

$$\tilde{O}\left(dT^{\frac{1-\epsilon}{2(1+\epsilon)}}\sqrt{\sum_{t=1}^{T}\nu_t^2}+dT^{\frac{1-\epsilon}{2(1+\epsilon)}}\right)$$

- Worst-case regret:

$$\tilde{O}\left(dT^{\frac{1}{1+\epsilon}}\right),\Omega\left(dT^{\frac{1}{1+\epsilon}}\right)$$

- Deterministic-case regret:

$$\tilde{O}(d),\Omega(d)$$

Linear MDPs

- adaptive Huber regression
- weighted ridge regression
- value iteration scheme
- rare-switching policy

Compute $\boldsymbol{\theta}_{k-1,h}$, $\widehat{\boldsymbol{w}}_{k-1,h}$ and $\check{\boldsymbol{w}}_{k-1,h}$ via (6.1) and (6.8).

if UPDATE **then**

$$Q_h^k(\cdot,\cdot) = \langle \boldsymbol{\phi}(\cdot,\cdot), \boldsymbol{\theta}_{k-1,h} + \widehat{\boldsymbol{w}}_{k-1,h} \rangle + \beta_{R,k-1} \|\boldsymbol{\phi}(\cdot,\cdot)\|_{\boldsymbol{H}_{k-1,h}^{-1}} + \beta_V \|\boldsymbol{\phi}(\cdot,\cdot)\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}.$$

$$\check{Q}_h^k(\cdot,\cdot) = \langle \boldsymbol{\phi}(\cdot,\cdot), \boldsymbol{\theta}_{k-1,h} + \check{\boldsymbol{w}}_{k-1,h} \rangle - \beta_{R,k-1} \|\boldsymbol{\phi}(\cdot,\cdot)\|_{\boldsymbol{H}_{k-1,h}^{-1}} - \beta_V \|\boldsymbol{\phi}(\cdot,\cdot)\|_{\boldsymbol{\Sigma}_{k-1,h}^{-1}}.$$

$$Q_h^k(\cdot,\cdot) = \min\{Q_h^k(\cdot,\cdot), Q_h^{k-1}(\cdot,\cdot), \mathcal{H}\}, \check{Q}_h^k(\cdot,\cdot) = \max\{\check{Q}_h^k(\cdot,\cdot), \check{Q}_h^{k-1}(\cdot,\cdot), 0\}.$$

Set $k_{\text{last}} = k$.

else

$$Q_h^k(\cdot,\cdot) = Q_h^{k-1}(\cdot,\cdot),\ \check{Q}_h^k(\cdot,\cdot) = \check{Q}_h^{k-1}(\cdot,\cdot).$$

end if

$$V_h^k(\cdot) = \max_a Q_h^k(\cdot,a),\ \check{V}_h^k(\cdot) = \max_a \check{Q}_h^k(\cdot,a),\ \pi_h^k(\cdot) = \operatorname{argmax}_a Q_h^k(\cdot,a).$$

Result of Linear MDPs

- Regret:

$$\tilde{O}\left(d\sqrt{H\mathcal{U}^*}K^{\frac{1}{1+\epsilon}} + d\sqrt{H\mathcal{V}^*K}\right)$$

- First-order regret:

$$\tilde{O}\left(d\sqrt{H^2V_1^*K}\right)$$

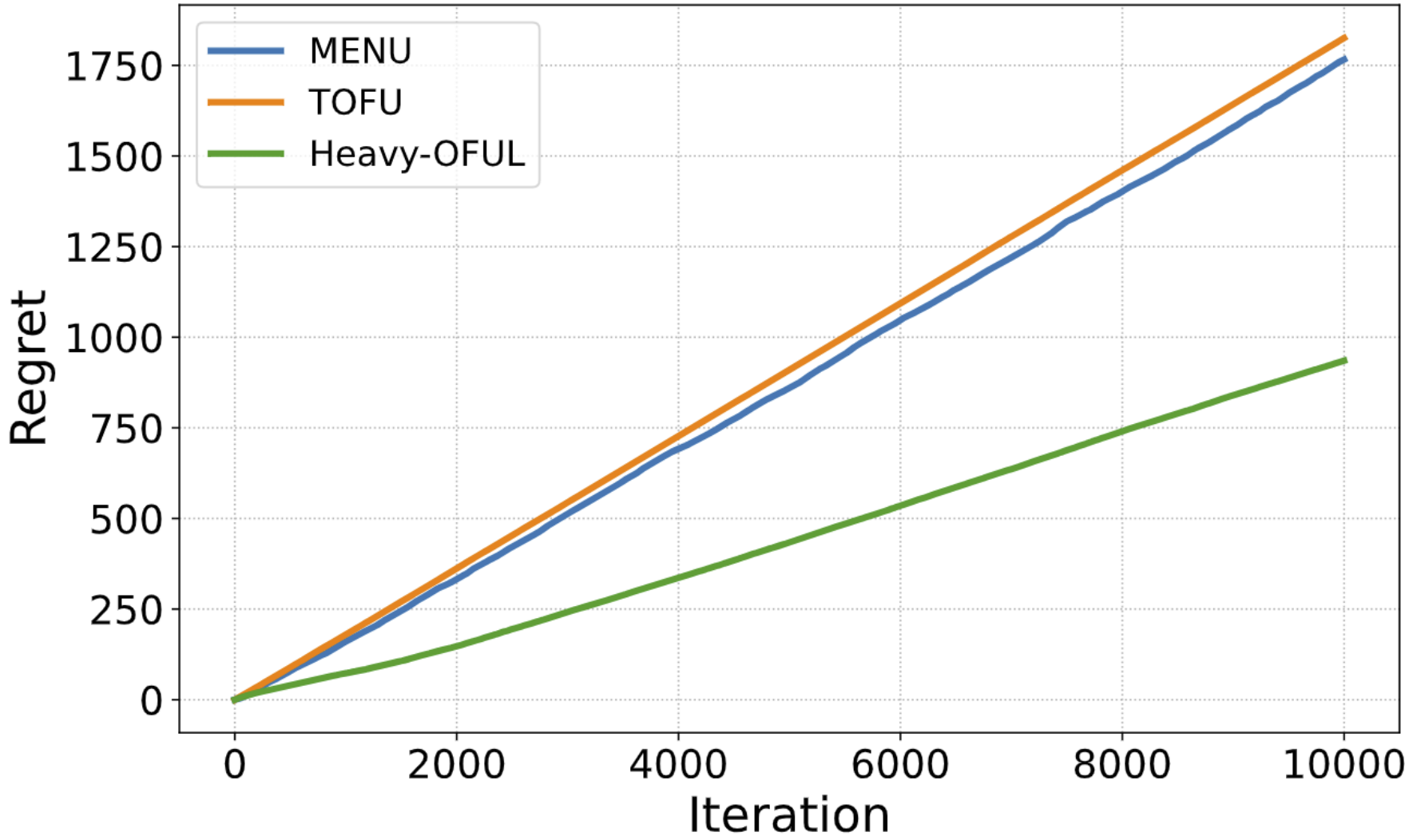
- Worst-case regret:

$$\tilde{O}\left(dHK^{\frac{1}{1+\epsilon}} + d\sqrt{H^3K}\right)$$

- Lower bound:

$$\Omega\left(dHK^{\frac{1}{1+\epsilon}} + d\sqrt{H^3K}\right)$$

Experimental Result



Thanks for listening!