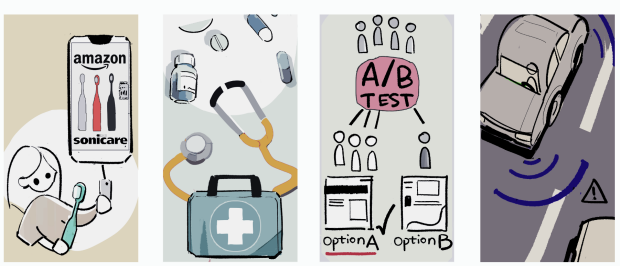
Online Estimation and Inference for Robust Policy Evaluation in Reinforcement Learning

Weidong Liu, Jiyuan Tu, Xi Chen, Yichen Zhang

NeurIPS 2025

Sequential Decision-Making



- Regret Minimization
  - Bandit/RL algorithms are designed for regret minimization
  - Regret: how much worse it performs compared to an offline oracle
- Statistical inference or uncertainty quantification
  - Infer the effect of a policy; Gaining generalizable knowledge; Identifying new directions; Crucial for scientific discovery

Statistical Inference in Sequential Decision-Making

- Bandit /RL Algorithms Induce Dependence
  - Observations are not independent over time
    - Use past history to select the action in the new context
    - Data is adaptively collected
  - Consequences: Introduce a bias, or potentially non-normality
    - Villar, Bowden, Wason (2015); Deshpande et al (2021); Khamaru, Mackey, Wainwright (2021); Zhang, Janson, Murphy (2021, 2022); Chen, Lu, Song (2021ab)
    - Not fully-online algorithms: Heavy cost on computation and storage
- Design inference procedures that are simultaneously: sample-efficient, computation-efficient, valid for adaptive collected data?

Policy Evaluation w. Linear Function Approximation

For an MDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$, we evaluate the cumulative reward for a policy π

$$J^*(s) = \mathbb{E}\Big[\sum_{k=0}^{\infty} \gamma^k \mathcal{R}(s_k) \;\Big|\; s_0 = s\Big].$$

- Linear function approximation: $\widetilde{J}(s, \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \boldsymbol{\phi}(s) = \sum_{l=1}^{d} \theta_l \phi_l(s)$.
- The optimal parameter minimizes the MSPBE (Tsitsiklis and Van Roy, 1997):

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta}} \left\| \widetilde{\boldsymbol{J}}(\boldsymbol{\theta}) - \Pi \mathcal{T}^\pi \widetilde{\boldsymbol{J}}(\boldsymbol{\theta}) \right\|_D^2,$$

  where D denotes the stable distribution under π, and $\mathcal{T}$ is the Bellman operator:

$$\mathcal{T}(Q) = \mathcal{R} + \gamma \mathcal{P} Q.$$

LEAST-SQUARES TD(0)

- Consider an equivalent fixed-point formulation (Kolter and Ng, 2009):

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^d} \mathbb{E}|\boldsymbol{\phi}^\top(s)\boldsymbol{u} - (\mathcal{R}(s) + \gamma\boldsymbol{\phi}^\top(s')\boldsymbol{\theta}^*)|^2.$$

- Given a trajectory $\{(s_n, \mathcal{R}(s_n), s_{n+1})\}_{n=1}^{\infty}$, the linear TD update (Sutton, 1988):

$$\widehat{\boldsymbol{\theta}}_{n+1} = \widehat{\boldsymbol{\theta}}_n + \alpha_{n+1}\big\{(\boldsymbol{\phi}(s_n) - \gamma\boldsymbol{\phi}^\top(s_{n+1}))^\top\widehat{\boldsymbol{\theta}}_n - \mathcal{R}(s_n)\big\}\boldsymbol{\phi}(s_n).$$

- Problems:
  - Linear TD(0) algorithm is vulnerable to corruption in reward $\mathcal{R}$;
  - How to conduct online inference on $\boldsymbol{\theta}^*$, or $J^*(s)$?

Robust Temporal Difference

- Reward $\mathcal{R}$ comes from distribution $(1-\alpha_n)P + \alpha_n Q$:
  - Q denotes arbitrary outlier distribution;
  - P has $(1+\delta)$-moment;
- Pseudo Huber loss $f_\tau(x) = \tau^2(\sqrt{1+(x/\tau)^2} - 1)$;
  - Smoothed version of Huber: near quadratic locally, Lipschitz globally.
- Solve the robust fixed-point equation

$$\boldsymbol{\theta}_\tau^* = \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^d} \mathbb{E}\, f_\tau\big(\boldsymbol{\phi}^\top(s)\boldsymbol{u} - (\mathcal{R}(s) + \gamma \boldsymbol{\phi}^\top(s')\boldsymbol{\theta}_\tau^*)\big);$$

- No longer quadratic $\Rightarrow$ Gradient methods performs nonlinear updates
- Equivalent to $\mathbb{E}\big[\boldsymbol{\phi}(s) g_\tau\big((\boldsymbol{\phi}^\top(s) - \gamma\boldsymbol{\phi}^\top(s'))\boldsymbol{\theta}_\tau^* - \mathcal{R}(s)\big)\big] = 0$, with $g_\tau(x) = f_\tau'(x)$.

Online Newton Method

- Denote $\boldsymbol{X}_i = \boldsymbol{\phi}(s_i)$, $\boldsymbol{Z}_i = \boldsymbol{\phi}(s_i) - \gamma\boldsymbol{\phi}(s_{i+1})$, and $b_i = \mathcal{R}(s_i)$;
- Target problem can be written as

$$\mathbb{E}\big[\boldsymbol{X} g_\tau(\boldsymbol{Z}^\top \boldsymbol{\theta}_\tau^* - b)\big] = 0;$$

- Propose the following online Newton update:

$$\widehat{\boldsymbol{\theta}}_{n+1} = \frac{1}{n+1}\sum_{i=0}^{n}\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{H}}_{n+1}^{-1}\frac{1}{n+1}\sum_{i=0}^{n}\boldsymbol{X}_{i+1} g_{\tau_{i+1}}(\boldsymbol{Z}_{i+1}^\top\widehat{\boldsymbol{\theta}}_i - b_{i+1}),$$

$$\text{with}\quad \widehat{\boldsymbol{H}}_{n+1} = \frac{1}{n+1}\sum_{i=0}^{n}\boldsymbol{X}_{i+1}\boldsymbol{Z}_{i+1}^\top g'_{\tau_{i+1}}(\boldsymbol{Z}_{i+1}^\top\widehat{\boldsymbol{\theta}}_i - b_{i+1}).$$

Online Newton Method

- Reformulate the update as

$$\widehat{\boldsymbol{\theta}}_{n+1} = \bar{\boldsymbol{\theta}}_{n+1} - \widehat{\boldsymbol{H}}_{n+1}^{-1}\boldsymbol{G}_{n+1};$$

- Online gradient update

$$\boldsymbol{G}_{n+1} = \frac{n}{n+1}\boldsymbol{G}_n + \frac{1}{n+1}\boldsymbol{X}_{n+1}g_{\tau_{n+1}}(\boldsymbol{Z}_{n+1}^\top \widehat{\boldsymbol{\theta}}_n - b_{n+1});$$

- Online Hessian update (by Sherman-Morrison formula) – only scalar inverse:

$$\begin{aligned}\widehat{\boldsymbol{H}}_{n+1}^{-1} =& \frac{n+1}{n}\widehat{\boldsymbol{H}}_n^{-1} - \frac{n+1}{n^2}\widehat{\boldsymbol{H}}_n^{-1}\boldsymbol{X}_{n+1} \\ & \times \left[\frac{1}{n}\boldsymbol{Z}_{n+1}^\top \widehat{\boldsymbol{H}}_n^{-1}\boldsymbol{X}_{n+1} + \{g'_{\tau_{n+1}}(\boldsymbol{Z}_{n+1}^\top \widehat{\boldsymbol{\theta}}_n - b_{n+1})\}^{-1}\right]^{-1}\boldsymbol{Z}_{n+1}^\top \widehat{\boldsymbol{H}}_n^{-1}.\end{aligned}$$

Convergence Rate

Theorem 1

Take the thresholding parameter $\tau_i = C_\tau i^\beta$. Assume n_0 is sufficiently large and the initial value $|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*|_2 \le c_0$ for some $c_0 < 1$. Then there is

$$\mathbb{P}\Big(\cap_{i=n_0}^{n} \big\{ |\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}^*|_2 \ge C\sqrt{d} e_i \big\} \Big) \ge 1 - c n_0^{-\nu},$$

where $$e_n = \alpha_n \tau_n + \tau_n^{-\min(\delta,2)} + \sqrt{\frac{\tau_n^{(1-\delta)_+} \log n}{n}} + \frac{\tau_n \log^2 n}{n} + \frac{1}{\sqrt{d}} (c_0)^{2^{n-n_0}}.$$

Assume $m_n = \alpha_n n$ outliers among n samples. We further have

$$|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*|_2 = O_{\mathbb{P}}\left(\sqrt{d} \Big(\sqrt{\frac{\log n}{n}} + \frac{\log^2 n}{n^{1-\beta}} + \frac{m_n}{n^{1-\beta}} + \frac{1}{n^{2\beta}} \Big) \right).$$

Convergence Rate (No Contamination)

Corollary 2

When the contamination rate $\alpha_n = 0$,

- When $\delta \in (0, 1]$, we specify $\tau_i = C_\tau (i/\log i)^{1/(1+\delta)}$. Then

$$|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*|_2 = O_{\mathbb{P}}\left(\sqrt{d}\left(\frac{\log n}{n}\right)^{\frac{\delta}{1+\delta}}\right).$$

- When $\delta > 1$, we specify $\tau_i = C_\tau (i/\log i)^{1/2}$. Then

$$|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*|_2 = O_{\mathbb{P}}\left(\sqrt{d}\left(\frac{\log n}{n}\right)^{1/2}\right).$$

Asymptotic Normality

Theorem 3

If the contamination rate $\alpha_n = o(1/(\sqrt{n}\tau_n))$ and $n^{1/4} = o(\tau_n)$,

$$\frac{\sqrt{n}}{\sigma_{\boldsymbol{v}}} \boldsymbol{v}^\top (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(0,1), \text{ where } \sigma_{\boldsymbol{v}}^2 = \boldsymbol{v}^\top \boldsymbol{H}^{-1} \boldsymbol{\Sigma} (\boldsymbol{H}^\top)^{-1} \boldsymbol{v},$$

as $n \to \infty$. Here $\boldsymbol{\Sigma} = \sum_{k=-\infty}^{\infty} \mathbb{E}\big[\boldsymbol{X}_0 \boldsymbol{X}_k^\top (\boldsymbol{Z}_0^\top \boldsymbol{\theta}^* - b_0)(\boldsymbol{Z}_k^\top \boldsymbol{\theta}^* - b_k)\big]$.
When $\alpha_n = 0$ (no contamination) and $\tau_i = C_\tau i^\beta$ for $\beta \geq 3/4$,

$$\sqrt{n} \boldsymbol{v}^\top (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) = \boldsymbol{v}^\top \boldsymbol{H}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{X}_i (\boldsymbol{Z}_i^\top \boldsymbol{\theta}^* - b_i) + O_{\mathbb{P}}\Big(\frac{d \log n}{\sqrt{n}}\Big).$$

- Asymptotic Normality is valid when $d^2/n = o(1)$, ignoring logarithm terms.
- For first-order methods, best remainder rate is $O_{\mathbb{P}}(dn^{-1/3})$, even under i.i.d.

Online Statistical Inference

- Asymptotic covariance matrix $\boldsymbol{H}^{-1}\boldsymbol{\Sigma}(\boldsymbol{H}^\top)^{-1}$ with

$$\boldsymbol{\Sigma} = \sum_{k=-\infty}^{\infty} \mathbb{E}\big[\boldsymbol{X}_0\boldsymbol{X}_k^\top(\boldsymbol{Z}_0^\top\boldsymbol{\theta}^* - b_0)(\boldsymbol{Z}_k^\top\boldsymbol{\theta}^* - b_k)\big].$$

- Construct the estimator

$$\widehat{\boldsymbol{\Sigma}}_n = \frac{1}{n}\sum_{i=1}^{n} \sum_{k=-\lceil\lambda\log i\rceil\wedge(i-1)}^{\lceil\lambda\log i\rceil\wedge(i-1)} \boldsymbol{X}_i\boldsymbol{X}_{i-k}^\top g_{\tau_i}(\boldsymbol{Z}_i^\top\widehat{\boldsymbol{\theta}}_{i-1} - b_i) g_{\tau_{i-k}}(\boldsymbol{Z}_{i-k}^\top\widehat{\boldsymbol{\theta}}_{i-k-1} - b_{i-k}).$$

- Requires only $O(\lceil\log n\rceil)$ memory due to ϕ-mixing property.

Theorem 4

The covariance estimator $\widehat{\boldsymbol{\Sigma}}_n$ satisfies

$$\|\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\| = O_{\mathbb{P}}\left(\sqrt{d}\tau_n^2\alpha_n + \sqrt{d}\tau_n^{-1} + \tau_n\sqrt{\frac{d\log n}{n}} + \frac{d\tau_n^2\log^2 n}{n}\right).$$

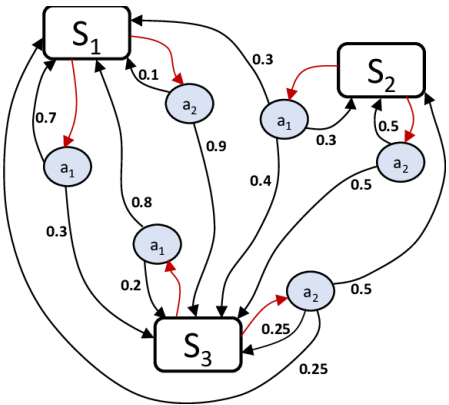
Given a unit vector $\boldsymbol{v}$, a confidence interval with nominal level $(1-\xi)$ is

$$\left[\boldsymbol{v}^\top\widehat{\boldsymbol{\theta}}_n - q_{1-\xi/2}\widehat{\sigma}_{\boldsymbol{v}}, \boldsymbol{v}^\top\widehat{\boldsymbol{\theta}}_n + q_{1-\xi/2}\widehat{\sigma}_{\boldsymbol{v}}\right],$$

with $\widehat{\sigma}_{\boldsymbol{v}}^2 = \boldsymbol{v}^\top\widehat{\boldsymbol{H}}_n^{-1}\widehat{\boldsymbol{\Sigma}}_n(\widehat{\boldsymbol{H}}_n^\top)^{-1}\boldsymbol{v}$ and $q_{1-\xi/2} = \Phi^{-1}(1-\xi/2)$.

Simulation Studies

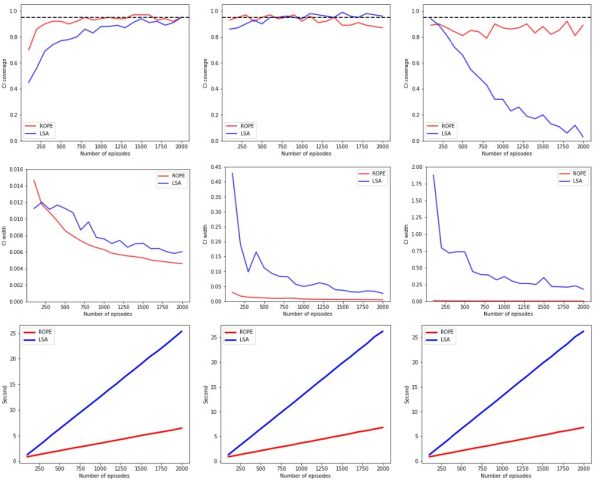
Infinite-Horizon MDP



FrozenLake Environment

FrozenLake Environment

Comparative Analysis



- Comparison between ROPE and LSA;
- Coverage probability (the first row)
  the width of confidence interval (the second row)
  computing time (the third row) ;
- Contamination rate α_n in $\{0, n^{-1}, 0.05n^{-1/2}\}$.

Conclusions

- An online policy evaluation robust to heavy-tailedness and outliers;
- Technically difficult due to non-linear stochastic gradient update;
- An online statistical inference procedure for policy evaluation;
- 2nd-order approach for improved efficiency without computation loss;
- Improved higher-order convergence to the asymptotic normality.
- The *Annals of Statistics* (to appear), arXiv: 2310.02581.

Thanks!