A Finite Sample Complexity Bound for Distributionally Robust Q-learning

Shengbo Wang*, Nian Si†, Jose Blanchet*, Zhengyuan Zhou‡

shengbo.wang@stanford.edu, niansi@chicagobooth.edu, jose.blanchet@stanford.edu, zz26@stern.nyu.edu

*Stanford MS&E, †Chicago Booth, ‡NYU Stern

Contributions

1. Extend the Multilevel Monte Carlo (MLMC) based distributionally robust (DR) Bellman estimator in Liu et al. (2022) such that the expected sample size of constructing our estimator is of *constant order*.
2. Establish the MLMC DR Q-learning algorithm and prove that the expected sample complexity of our algorithm is $\tilde{O}\left(|S||A|(1-\gamma)^{-5}\epsilon^{-2}p_\wedge^{-6}\delta^{-4}\right)$. This is tight in $|S||A|$ and nearly tight in the effective horizon $(1-\gamma)^{-1}$ at the same time.
3. The first model-free algorithm and analysis that guarantee solving the DR-RL problem with a finite expected sample complexity.
4. Numerically exhibit the validity of our theorem predictions and demonstrate the improvements of our algorithm over that in Liu et al. (2022).

Distributionally Robust Markov Decision Processes

$\mathcal{M}_0 = (S, A, R, \mathcal{P}_0, \mathcal{R}_0, \gamma)$ an MDP, where S, A, and $R \subsetneq \mathbb{R}_{\geq 0}$ are finite state, action, and reward spaces. $\mathcal{P}_0 = \{p_{s,a}, s \in S, a \in A\}$ and $\mathcal{R}_0 = \{\nu_{s,a}, s \in S, a \in A\}$ are the sets of the reward and transition distributions. KL uncertainty sets $\mathcal{P}_{s,a}(\delta) \coloneqq \{p : D_{\mathrm{KL}}(p \| p_{s,a}) \leq \delta\}$ and $\mathcal{R}_{s,a}(\delta) \coloneqq \{\nu : D_{\mathrm{KL}}(\nu \| \nu_{s,a}) \leq \delta\}$.

- Min-max control problem for history dependent controller and adversary

$$V^*(s) = \sup_{\pi \in \Pi} \inf_{P \in \mathcal{K}^\pi(\delta)} \mathbb{E}_P\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| s_0 = s\right]$$

- Markov optimality: There exists a Markovian policy that is optimal to the min-max control. Under this policy, the optimal adversarial distribution choice is Markovian as well.
- The Distributionally robust optimal Q-function and its Bellman equation

$$\begin{aligned} Q^*(s,a) &\coloneqq \mathbb{E}_{r\sim\nu_{s,a}}[r] + \gamma \mathbb{E}_{s'\sim p_{s,a}}[V^*(s')] \\ &= \mathbb{E}_{r\sim\nu_{s,a}}[r] + \gamma \mathbb{E}_{s'\sim p_{s,a}}\left[\max_{b\in A} Q^*(s',b)\right] \\ &=: \mathcal{T}_\delta(Q^*). \end{aligned}$$

- Optimal policy: $\pi^*(s) = \arg\max_{a\in A} Q^*(s,a)$.

Strong Duality

Hu and Hong (2013), Theorem 1.

$$\sup_{P:D_{KL}(P\|P_0)\leq\delta} \mathbb{E}_P[H(X)] = \inf_{\alpha\geq 0}\left\{\alpha \log \mathbb{E}_{P_0}\left[e^{H(X)/\alpha}\right] + \alpha\delta\right\}.$$

Dual Formulation of DR-RL Problem

The *dual form* of the DR Bellman Operator

$$\mathcal{T}_\delta(Q)(s,a) = \sup_{\alpha\geq 0}\left\{-\alpha \log \mathbb{E}_{r\sim\nu_{s,a}}\left[e^{-r/\alpha}\right] - \alpha\delta\right\} + \gamma \sup_{\beta\geq 0}\left\{-\beta \log \mathbb{E}_{s'\sim p_{s,a}}\left[e^{-v(Q)(s')/\beta}\right] - \beta\delta\right\}.$$

Learn the unique solution Q^* of the fixed point equation $\mathcal{T}(Q) = Q$ using samples from $\mathcal{P}_0$ and $\mathcal{R}_0$.

Multilevel Monte Carlo DR Bellman Operator

For given $g \in (0,1)$ and $Q \in \mathbb{R}^{S\times A}$, define the *MLMC-DR estimator*:

$$\widehat{\mathcal{T}}_{\delta,g}(Q)(s,a) \coloneqq \widehat{R}_\delta(s,a) + \gamma \widehat{V}_\delta(Q)(s,a).$$

For $\widehat{R}_\delta(s,a)$ and $\widehat{V}_\delta(s,a)$, we sample N_1, N_2 from a geometric distribution $\mathrm{Geo}(g)$. Draw 2^{N_1+1} samples $r_i \sim \nu_{s,a}$ and 2^{N_2+1} samples $s_i' \sim p_{s,a}$. Compute

$$\widehat{R}_\delta(s,a) \coloneqq r_1 + \frac{\Delta^R_{N_1,\delta}}{p_{N_1}}, \qquad \widehat{V}_\delta(Q)(s,a) \coloneqq v(Q)(s_1') + \frac{\Delta^P_{N_2,\delta}(Q)}{p_{N_2}}$$

where

$$\begin{aligned} \Delta^R_{n,\delta} = &\sup_{\alpha\geq 0}\left\{-\alpha \log \mathbb{E}_{r\sim\nu_{s,a,2^{n+1}}}\left[e^{-r/\alpha}\right] - \alpha\delta\right\} \\ &- \frac{1}{2}\sup_{\alpha\geq 0}\left\{\alpha \log \mathbb{E}_{r\sim\nu^E_{s,a,2^n}}\left[e^{-r/\alpha}\right] - \alpha\delta\right\} - \frac{1}{2}\sup_{\alpha\geq 0}\left\{-\alpha \log \mathbb{E}_{r\sim\nu^O_{s,a,2^n}}\left[e^{-r/\alpha}\right] - \alpha\delta\right\} \end{aligned}$$

and

$$\begin{aligned} \Delta^P_{n,\delta}(Q) = &\sup_{\beta\geq 0}\left\{-\beta \log \mathbb{E}_{s'\sim p_{s,a,2^{n+1}}}\left[e^{-v(Q)(s')/\beta}\right] - \beta\delta\right\} \\ &- \frac{1}{2}\sup_{\beta\geq 0}\left\{-\beta \log \mathbb{E}_{s'\sim p^E_{s,a,2^n}}\left[e^{-v(Q)(s')/\beta}\right] - \beta\delta\right\} - \frac{1}{2}\sup_{\beta\geq 0}\left\{-\beta \log \mathbb{E}_{s'\sim p^O_{s,a,2^n}}\left[e^{-v(Q)(s')/\beta}\right] - \beta\delta\right\}. \end{aligned}$$

Properties of the MLMC-DR estimator:

- $\widehat{\mathcal{T}}_{\delta,g}$ is unbounded.
- $\widehat{\mathcal{T}}_{\delta,g}$ is *unbiased* for $\mathcal{T}_\delta$ for any δ, g; i.e. for any Q, $\mathbb{E}\widehat{\mathcal{T}}_{\delta,g}(Q) = \mathcal{T}_\delta(Q)$.
- Define $p_\wedge$ to be the minimum positive probability of $\mathcal{P}_0$ and $\mathcal{R}_0$. Assume $\delta = O(p_\wedge)$, then

$$\mathbb{E}\|\widehat{\mathcal{T}}_{\delta,g}(Q) - \mathcal{T}_\delta(Q)\|_\infty^2 \leq \tilde{O}\left(\frac{r_{\max}^2 + \gamma^2\|Q\|_\infty^2}{\delta^4 p_\wedge^6}\right).$$

MLMC DR Q-Learning

The MLMC DR Q-Learning algorithm:

- Input step size $\{\alpha_t\}$ and $g \in (0, 3/4)$.
- At each iteration k, sample independent MLMC DR Bellman operator $\widehat{\mathcal{T}}_{\delta,g,k+1}$ defined before.
- Perfore Q-Learning update

$$\widehat{Q}_{\delta,k+1} = (1-\alpha_t)\widehat{Q}_{\delta,k} + \alpha_k \widehat{\mathcal{T}}_{\delta,g,k+1}(\widehat{Q}_{\delta,k}).$$

Convergence Rates and Sample Complexities

Running the MLMC DR Q-learning until iteration k. The following holds:

- Constant step size: Choose

$$\alpha_k \equiv \alpha \leq \frac{(1-\gamma)^2\delta^4 p_\wedge^6}{c'\gamma^2\tilde{l}\log(|S||A|)},$$

  then we have

$$\mathbb{E}\|\widehat{Q}_{\delta,k} - Q^*_\delta\|_\infty^2 \leq \frac{3r_{\max}^2}{2(1-\gamma)^2}\left(1 - \frac{(1-\gamma)\alpha}{2}\right)^k + \frac{c\alpha r_{\max}^2 \log(|S||A|)\tilde{l}}{\delta^4 p_\wedge^6 (1-\gamma)^4}.$$

- Rescaled linear step size: Choose

$$\alpha_k = \frac{4}{(1-\gamma)(k+K)}, \qquad K = \frac{c'\tilde{l}\log(|S||A|)}{\delta^4 p_\wedge^6 (1-\gamma)^3},$$

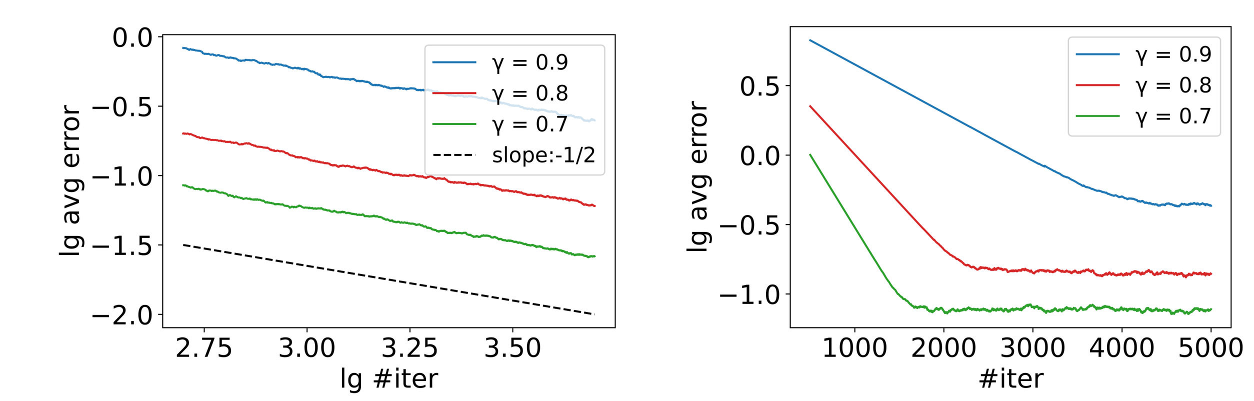
  then we have

$$\mathbb{E}\|\widehat{Q}_{\delta,k} - Q^*_\delta\|_\infty^2 \leq \frac{c r_{\max}^2 \tilde{l}\log(|S||A|)\log(k+K)}{\delta^4 p_\wedge^6 (1-\gamma)^5 (k+K)}.$$

The sample complexity under both step size to achieve ϵ error is

$$\tilde{O}\left(\frac{r_{\max}^2|S||A|}{\delta^4 p_\wedge^6 (1-\gamma)^5\epsilon^2}\right)$$

Numerical Results



(a) Rescaled linear step size. (b) Constant step size

Figure 1: Convergence of the MLMC DR Q-learning under the rescaled linear and constant step size. (a) shows lines with slop $-1/2$ which correspond to the $O(k^{-1/2})$ convergence rate. (b) shows geometric convergence initially and stays at constant error.

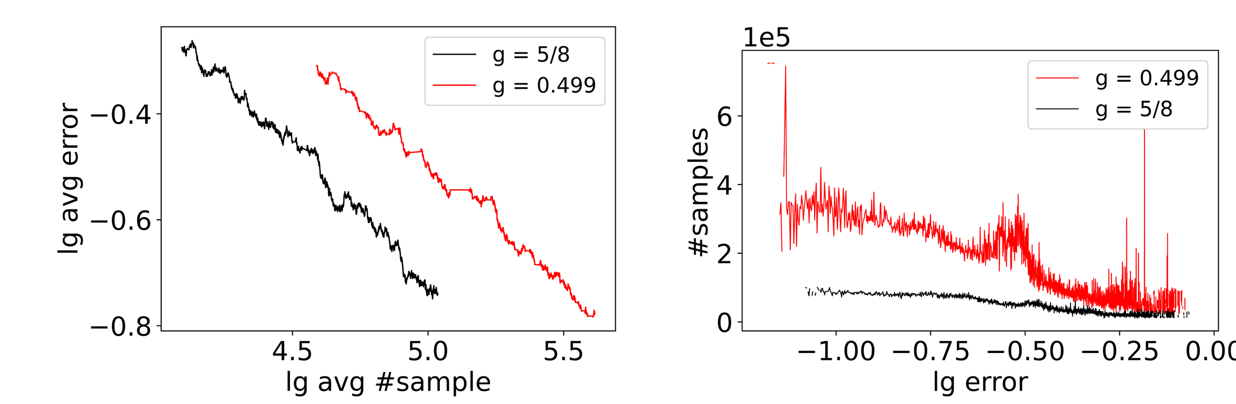


Figure 2: Performance comparison of our algorithm (black) and that in Liu et al. Our algorithm achieve better error with less samples