Polyak-Ruppert-Averaged Q-Learning is Statistically Efficient

Xiang Li
Joint work with Wenhao Yang, Jiadong Liang, Zhihua Zhang, Michael I. Jordan

Peking University, University of California Berkeley

July 11, 2022

Markov Decision Process (MDP)

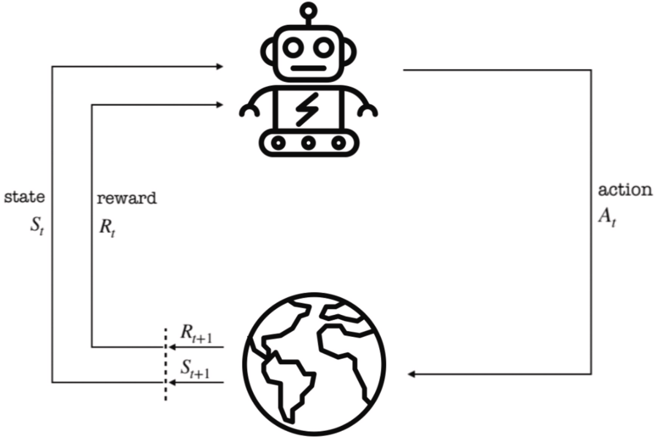

Figure 1: Illustration of a MDP [Perera at al., 2021]

Discounted infinite-horizon MDPs

- An infinite-horizon MDP is represented by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \gamma, P, R, r)$ with the state space $\mathcal{S}$, the action space $\mathcal{A}$ and the discount factor $\gamma \in [0, 1)$.
- $P \colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ represents the probability transition kernel.
- $R \colon \mathcal{S} \times \mathcal{A} \to [0, \infty)$ stands for the random reward and $r = \mathbb{E}R$.
- A policy $\pi : \mathcal{S} \to \mathcal{A}$ and its (Q-)value is defined to be

$$V^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s \right]$$

$$Q^\pi(s, a) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s, a_0 = a \right]$$

- Target: find the optimal policy $\pi^*(s) = \operatorname{argmax}_\pi V^\pi(s)$ and its value function $V^* := V^{\pi^*}$ and $Q^* := Q^{\pi^*}$.

Q-learning

- Only need to solve the Bellman equation: for $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$Q^*(s, a) = r(s, a) + \gamma \mathcal{T}(Q^*)(s, a)$$

$$\mathcal{T}(Q)(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a' \in \mathcal{A}} Q(s', a'). \tag{1}$$

- Q-learning [Watkins, 1989] is perhaps the most popular model-free learning algorithm in RL.

$$Q_t = (1 - \eta_t) Q_{t-1} + \eta_t \widehat{\mathcal{T}}_t(Q_{t-1}) \text{ where} \tag{2}$$

- $\widehat{\mathcal{T}}_t$ is an independent estimate of $\mathcal{T}$:

$$\widehat{\mathcal{T}}_t(Q)(s, a) = r_t(s, a) + \gamma \max_{a' \in \mathcal{A}} Q(s_t, a'), \tag{3}$$

with $r_t(s, a) \sim R(s, a)$ and $s_t = s_t(s, a) \sim P(\cdot|s, a)$ for each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Sample Complexity of Q-learning

- Sample efficiency = # of samples to achieve ε-accuracy.
- Each generation of $\widehat{\mathcal{T}}_t$ require $O(|\mathcal{S} \times \mathcal{A}|)$ samples.
- The sample efficiency of Q-learning is $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}\times\mathcal{A}|}{(1-\gamma)^{4}\varepsilon^2}\right)$ tight up to a log factor [Li et al., 2021, 2020].
- The minimax lower bound is $\Omega\left(\frac{|\mathcal{S}\times\mathcal{A}|}{(1-\gamma)^{3}\varepsilon^2}\right)$ [Azar et al., 2013].

Question

Could we find a model-free method to close the gap on $(1-\gamma)^{-1}$.

Current Solutions

- Previous model-free works close the gap via variance reduction.
- Variance reduction uses the following re-centered operator

$$\widehat{\mathcal{T}}_t^{\mathrm{VR}}(Q) := \widehat{\mathcal{T}}_t(Q) - \widehat{\mathcal{T}}_t(\overline{Q}) + \widehat{\overline{\mathcal{T}}}(\overline{Q})$$

  with $\overline{Q}$ the estimation in last epoch and $\widehat{\overline{\mathcal{T}}}$ an independent empirical $\mathcal{T}$ using more data [Wainwright, 2019, Khamaru et al., 2021].
- $\overline{Q}$ and $\widehat{\overline{\mathcal{T}}}$ are updated less frequently and the overall sample complexity achieves the optimal.

Question

Could we find a simper variant of Q-learning to close the gap?

Averaged Q-learning

- The averaged iterates generated by a stochastic approximation (SA) algorithm has favorable asymptotic statistical properties Ruppert [1988] and Polyak and Juditsky [1992].
- The Polyak-Ruppert averaging of Q-learning is

$$\bar{Q}_T = \frac{1}{T}\sum_{t=1}^{T} Q_t$$

with $\{Q_t\}_{t\geq 0}$ updated as in Eq. (2).

- Use the averaged iterate $\bar{Q}_T$ rather than the last-iterate Q_T to do inference.
- Application in deep RL, benefits in error reduction and stability [Lillicrap et al., 2016, Anschel et al., 2017].

Matrix Notation

- Let $D = |\mathcal{S} \times \mathcal{A}|$. Denote the transition matrix $\boldsymbol{P} \in \mathbb{R}^{D \times S}$
- For a deterministic policy π, the introduced transition matrix by π is $\boldsymbol{P}^{\pi} := \boldsymbol{P}\boldsymbol{\Pi}^{\pi} \in \mathbb{R}^{D \times D}$ and $\boldsymbol{P}_{\pi} := \boldsymbol{\Pi}^{\pi}\boldsymbol{P} \in \mathbb{R}^{S \times S}$ where $\boldsymbol{e}_i$ the i-th standard basis vector and

$$\boldsymbol{\Pi}^{\pi} = \mathrm{diag}\{\boldsymbol{e}_{\pi(1)}^{\top}, \boldsymbol{e}_{\pi(2)}^{\top}, \cdots, \boldsymbol{e}_{\pi(S)}^{\top}\} \in \{0, 1\}^{S \times D}.$$

- The vector-form update rule is

$$\pi_{t-1} = \mathrm{greedy}(\boldsymbol{Q}_{t-1})$$
$$\boldsymbol{V}_{t-1} = \boldsymbol{\Pi}^{\pi_{t-1}}\boldsymbol{Q}_{t-1}$$
$$\boldsymbol{Q}_t = (1 - \eta_t)\boldsymbol{Q}_{t-1} + \eta_t(\boldsymbol{r}_t + \gamma \boldsymbol{P}_t \boldsymbol{V}_{t-1}).$$

Bellman Noise

- $\boldsymbol{Z}_t \in \mathbb{R}^D$ be the Bellman noise at the t-th iteration, whose (s, a)-th entry is

$$Z_t(s, a) = \widehat{\mathcal{T}}_t(Q^*)(s, a) - \mathcal{T}(Q^*)(s, a). \tag{4}$$

- Matrix form $\boldsymbol{Z}_t = (\boldsymbol{r}_t - \boldsymbol{r}) + \gamma(\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}^*$.
- An important quantity in our analysis is the covariance matrix of $\boldsymbol{Z}$

$$\mathrm{Var}(\boldsymbol{Z}) = \mathbb{E}_{r_t, s_t}\boldsymbol{Z}\boldsymbol{Z}^\top \in \mathbb{R}^{D \times D}. \tag{5}$$

Central Limit Theorem

Assumption

Assume (i) $0 \le \sup_{s,a} R(s,a) \le 1$*; (ii)* π^* *is unique; (iii)* $\eta_t = t^{-\alpha} (0.5 < \alpha < 1)$.

Theorem (Asymptotic normality for Q^*)

Under the assumption, we have

$$\sqrt{T}(\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*) \xrightarrow{d} \mathcal{N}(0, Var_{\boldsymbol{Q}}),$$

where the asymptotic variance is given by

$$Var_{\boldsymbol{Q}} = (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1} \mathrm{Var}(\boldsymbol{Z})(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-\top} \in \mathbb{R}^{D \times D}. \quad (6)$$

Here $\mathrm{Var}(\boldsymbol{Z})$ *is the covariance matrix of the Bellman noise* $\boldsymbol{Z}$ *defined in* (5).

Insights on Sample Efficiency

- By $\sqrt{T}(\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*) \xrightarrow{d} \mathcal{Z} \sim \mathcal{N}(0, \mathrm{Var}_{\boldsymbol{Q}})$ and the bounded convergence theorem,

$$\sqrt{T}\mathbb{E}\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty \to \mathbb{E}\|\mathcal{Z}\|_\infty \approx \sqrt{\ln D}\sqrt{\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty}.$$

- Requires about $T = \mathcal{O}\left(\frac{\ln D}{\varepsilon^2}\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty\right)$ iterations to ensure $\mathbb{E}\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty \le \varepsilon$.
- The difficulty indicator $\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty \le (1-\gamma)^{-3}$ [Azar et al., 2013, Khamaru et al., 2021]
- It seems averaged Q-learning could close the gap!

Question

Is the asymptotic variance optimal?

Semiparametric Efficiency Lower Bound

- The Cramer-Rao lower bound (CRLB) assesses the hardness of estimating a target parameter $\beta(\theta)$ in a parametric model $\mathcal{P}_\theta$ indexed by parameter θ.
- We meet semiparametric model here since the random reward $\{R(s,a)\}_{s,a}$ is fully nonparametric.
- Our MDP model $\mathcal{M}$ has parameter $\theta = (P, R)$.
- Denote the i.i.d. data we collected in T iterations is $\mathcal{D} = \{(\boldsymbol{r}_t, \boldsymbol{P}_t)\}_{t\in[T]}$.

Regular Asymptotically Linear (RAL) Estimator

Definition (Regular asymptotically linear)

We say that $\widehat{\boldsymbol{Q}}_T$ is regular asymptotically linear (RAL) for $\boldsymbol{Q}^*$ if it is regular and asymptotically linear with a measurable random function $\phi(\boldsymbol{r}_t, \boldsymbol{P}_t) \in \mathbb{R}^D$ such that

$$\sqrt{T}(\widehat{\boldsymbol{Q}}_T - \boldsymbol{Q}^*) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \phi(\boldsymbol{r}_t, \boldsymbol{P}_t) + o_{\mathbb{P}}(1),$$

where $\mathbb{E}\phi(\boldsymbol{r}_t, \boldsymbol{P}_t) = 0$ and $\mathbb{E}\phi(\boldsymbol{r}_t, \boldsymbol{P}_t)\phi(\boldsymbol{r}_t, \boldsymbol{P}_t)^\top$ is finite and nonsingular. Such a $\phi(\cdot, \cdot)$ is referred to as an influence function.

- An estimator is regular if its limiting distribution is unaffected by local changes in the data-generating process.

Regular Asymptotically Linear (RAL) Estimator

Theorem

Given the dataset $\mathcal{D} = \{(\boldsymbol{r}_t, \boldsymbol{P}_t)\}_{t\in[T]}$*, for any RAL estimator* $\widehat{\boldsymbol{Q}}_T$ *of* $\boldsymbol{Q}^*$ *computed from* $\mathcal{D} = \{(\boldsymbol{r}_t, \boldsymbol{P}_t)\}_{t\in[T]}$*, its variance satisfies*

$$\lim_{T\to\infty} T\mathbb{E}(\widehat{\boldsymbol{Q}}_T - \boldsymbol{Q}^*)(\widehat{\boldsymbol{Q}}_T - \boldsymbol{Q}^*)^\top \succeq Var_{\boldsymbol{Q}},$$

where $\boldsymbol{A} \succeq \boldsymbol{B}$ *means* $\boldsymbol{A} - \boldsymbol{B}$ *is positive semidefinite.*

Theorem

The averaged iterate $\bar{\boldsymbol{Q}}_T$ *is a RAL estimator for* $\boldsymbol{Q}^*$ *due to*

$$\sqrt{T}\left(\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\right) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1}\boldsymbol{Z}_t + o_{\mathbb{P}}(1).$$

Instance-dependent Convergence

Theorem

- *If* $\eta_t = t^{-\alpha}$ *with* $\alpha \in (0.5, 1)$, $\mathbb{E}\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty =$

$$\mathcal{O}\left(\sqrt{\|\mathrm{diag}(Var_{\boldsymbol{Q}})\|_\infty}\sqrt{\frac{\ln D}{T}} + \frac{\sqrt{\ln D}}{(1-\gamma)^3}\frac{1}{T^{1-\frac{\alpha}{2}}}\right) + \widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^{3+\frac{2}{1-\alpha}}}\frac{1}{T} + \frac{\gamma}{(1-\gamma)^{4+\frac{1}{1-\alpha}}}\frac{1}{T^\alpha}\right).$$

- *If* $\eta_t = \frac{1}{1+(1-\gamma)t}$, $\mathbb{E}\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty =$

$$\mathcal{O}\left(\sqrt{\frac{\|\mathrm{Var}(\boldsymbol{Z})\|_\infty}{(1-\gamma)^2}}\sqrt{\frac{\ln D}{T}}\right) + \widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^6}\frac{1}{T}\right).$$

Instance-dependent Convergence

- Match the instance optimality

$$\Omega\left(\sqrt{\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty}\sqrt{\frac{\ln D}{T}}\right).$$

- Sample complexity of variance-reduced Q-learning is

$$\mathcal{O}\left(\sqrt{\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty}\sqrt{\frac{\ln D}{T}}\right)+\widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^2}\frac{1}{T}\right).$$

- Not sure on whether linearly rescaled step sizes can match the lower bound since

$$\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty \le \frac{1}{(1-\gamma)^2}\|\mathrm{Var}(\boldsymbol{Z})\|_\infty$$

Instance-dependent Convergence

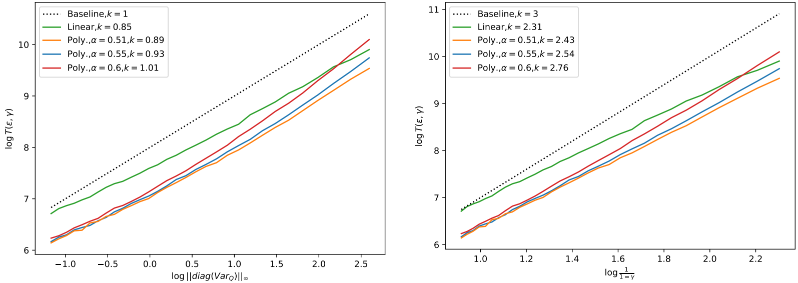

Figure 2: Log-log plots of the sample complexity $T(\varepsilon, \gamma)$ versus the asymptotic variance $\|\mathrm{diag}(\mathrm{Var}_{\boldsymbol{Q}})\|_\infty$ (left) and versus the discount complexity parameter $(1-\gamma)^{-1}$ (right).

Functional Central Limit Theorem (Donsker's Invariance Principle)

- With $\{X_t\}_{t\geq 0}$ i.i.d. r.v.'s with mean zero and unit variance and $S_n = \sum_{t=1}^{n} X_i$, the CLT yields $\frac{S_T}{\sqrt{T}} \xrightarrow{d} \mathcal{N}(0,1)$.
- Define $\phi_T(r) = \frac{S_{\lfloor Tr \rfloor}}{\sqrt{T}}$ with $r \in [0,1]$. The implies $\phi_T(r) \xrightarrow{w} \boldsymbol{B}_1(r)$ with $\boldsymbol{B}_1$ the 1-dim Brownian motion on $[0,1]$.

Functional Central Limit Theorem (FCLT)

In our case, define the standardized partial-sum processes as

$$\phi_T(r) := \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} (\boldsymbol{Q}_t - \boldsymbol{Q}^*), r \in [0, 1].$$

Theorem (FCLT)

Under the same assumptions,

$$\phi_T(\cdot) \xrightarrow{w} Var_{\boldsymbol{Q}}{}^{1/2} \boldsymbol{B}_D(\cdot), \tag{7}$$

where $Var_{\boldsymbol{Q}}$ is defined in (6) and $\boldsymbol{B}_D(\cdot)$ is the standard D-dimensional Brownian motion on $[0, 1]$. That is, for any given integer $n \geq 1$ and any $0 \leq t_1 < \cdots < t_n \leq 1$,

$$(\phi_T(t_1), \cdots, \phi_T(t_n)) \xrightarrow{d} Var_{\boldsymbol{Q}}{}^{1/2} (\boldsymbol{B}_D(t_1), \cdots, \boldsymbol{B}_D(t_n)).$$

Functional Central Limit Theorem (FCLT)

- By continuous mapping theorem, for any functional f on Càdlàg functions,

$$f(\phi_T) \xrightarrow{d} f(\mathrm{Var}_{\boldsymbol{Q}}{}^{1/2}\boldsymbol{B}_D(\cdot)).$$

- Construct a asymptotic pivotal statistic for inference.

Proposition

Letting f be a self-standalization function, we have

$$\phi_T(1)^\top \left(\int_0^1 \phi_T(r)\phi_T(r)^\top dr\right)^{-1} \phi_T(1) \xrightarrow{d} \boldsymbol{B}_D(1)^\top \left(\int_0^1 \boldsymbol{B}_D(r)\boldsymbol{B}_D(r)^\top dr\right)^{-1} \boldsymbol{B}_D(1).$$

FCLT for Statistic Inference

A close combination of optimization and statistics.

$$\phi_T(1)^\top \left(\int_0^1 \phi_T(r)\phi_T(r)^\top dr\right)^{-1} \phi_T(1)$$
$$\xrightarrow{d} \boldsymbol{B}_D(1)^\top \left(\int_0^1 \boldsymbol{B}_D(r)\boldsymbol{B}_D(r)^\top dr\right)^{-1} \boldsymbol{B}_D(1).$$

- The l.h.s. is a pivotal quantity involving samples and $\boldsymbol{Q}^*$.
- The pivotal quantity can computed fully.
- The r.h.s. is a known distribution (quantiles can be computed via simulation)
- Not the only choice of f.
- No need to estimate $\mathrm{Var}_{\boldsymbol{Q}}$ which is not easy.

Conclusion

- Averaged Q-learning is asymptotically optimal, achieving the established semeparametric Cramer-Rao lower bound.
- Averaged Q-learning achieves both the worst-case and instance-dependent optimality.
- We established a FCLT that helps conduct online statistical inference.

Thanks for listening!

References

Oron Anschel, Nir Baram, and Nahum Shimkin. Averaged-dqn: Variance reduction and stabilization for deep reinforcement learning. In *International Conference on Machine Learning*, pages 176–185. PMLR, 2017.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3):325–349, 2013.

Koulik Khamaru, Eric Xia, Martin J Wainwright, and Michael I Jordan. Instance-optimality in optimal value estimation: Adaptivity via variance-reduced Q-learning. *arXiv preprint arXiv:2106.14352*, 2021.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *arXiv preprint arXiv:2006.03041*, 2020.

Gen Li, Changxiao Cai, Yuxin Chen, Yuantao Gu, Yuting Wei, and Yuejie Chi. Is Q-learning minimax optimal? a tight sample complexity analysis. *arXiv preprint arXiv:2102.06548*, 2021.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *ICLR (Poster)*, 2016.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Martin J Wainwright. Variance-reduced Q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*, 2019.

Christopher Watkins. *Learning from delayed rewards*. PhD thesis, 1989.