A Statistical Analysis of Polyak-Ruppert Averaged Q-learning

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

Peking University, University of California Berkeley

Introduction

The classic Q-Learning:

$$Q_t = (1-\eta_t)Q_{t-1} + \eta_t \widehat{\mathcal{T}}_t(Q_{t-1}) \text{ where}$$
$$\widehat{\mathcal{T}}_t(Q)(s,a) = r_t(s,a) + \gamma \max_{a' \in \mathcal{A}} Q(s_t, a'),$$

is the empirical estimate of the Bellman operator

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

- Many analysis for the last-iterate Q_T, while the averaged iterate $\bar{Q}_T$ is less well understood,
$$\bar{Q}_T = \frac{1}{T}\sum_{t=1}^{T} Q_t.$$
- Classic results imply under mild regularity conditions,
$$\sqrt{T}(\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*) \xrightarrow{d} \mathcal{Z} \sim \mathcal{N}(\boldsymbol{0}, \mathsf{Var}_{\boldsymbol{Q}}),$$
where the asymptotic variance $\mathsf{Var}_{\boldsymbol{Q}}$ is given by
$$\mathsf{Var}_{\boldsymbol{Q}} := (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-1} \mathrm{Var}(\boldsymbol{Z}) (\boldsymbol{I} - \gamma \boldsymbol{P}^{\pi^*})^{-\top} \quad (1)$$
with $\boldsymbol{Z} \in \mathbb{R}^D$, $\boldsymbol{Z}(s,a) = [\widehat{\mathcal{T}}_t(Q) - \mathcal{T}(Q)](s,a)$.

Questions Studied

An observation from $\sqrt{T}(\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*) \xrightarrow{d} \mathcal{Z}$:

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

- How to obtain a valid non-asymptotic bound?
- Is this asymptotic variance $\mathsf{Var}_{\boldsymbol{Q}}$ optimal?
- Can we do statistical inference without estimating $\mathsf{Var}_{\boldsymbol{Q}}$?

Discounted infinite-horizon MDPs

- $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \gamma, P, R, r)$ with $\gamma \in [0,1)$.
- P: $\mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$: probability transition kernel.
- R: $\mathcal{S} \times \mathcal{A} \to [0, \infty)$: random reward, $r = \mathbb{E}R$.
- A policy $\pi : \mathcal{S} \to \mathcal{A}$ with its (Q-)value defined by
$$Q^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\Big| s_0 = s, a_0 = a\right].$$
- Assume unique optimal policy π^*.

Non-asymptotic Convergence

Assume $0 \le R(s,a) \le 1$ for all (s,a).

- If $\eta_t = t^{-\alpha}$ with $\alpha \in (0.5, 1)$, $\mathbb{E}\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty =$
$$\mathcal{O}\left(\sqrt{\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty}\sqrt{\frac{\ln D}{T}} + \frac{\sqrt{\ln D}}{(1-\gamma)^3}\frac{1}{T^{1-\frac{\alpha}{2}}}\right)$$
$$+\widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^{3+\frac{2}{1-\alpha}}}\frac{1}{T} + \frac{\gamma}{(1-\gamma)^{4+\frac{1}{1-\alpha}}}\frac{1}{T^\alpha}\right).$$
- The dominant term (in red) is of the same magnitude as the variance-reduced Q-learning.
- If $\eta_t = \frac{1}{1+(1-\gamma)t}$, $\mathbb{E}\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty =$
$$\mathcal{O}\left(\sqrt{\frac{\|\mathrm{Var}(\boldsymbol{Z})\|_\infty}{(1-\gamma)^2}}\sqrt{\frac{\ln D}{T}}\right) + \widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^6}\frac{1}{T}\right).$$
- Because $\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty \le \frac{1}{(1-\gamma)^2}\|\mathrm{Var}(\boldsymbol{Z})\|_\infty$, this rate is slightly lose. (How to tighten it? Unclear).

Numeral Validation

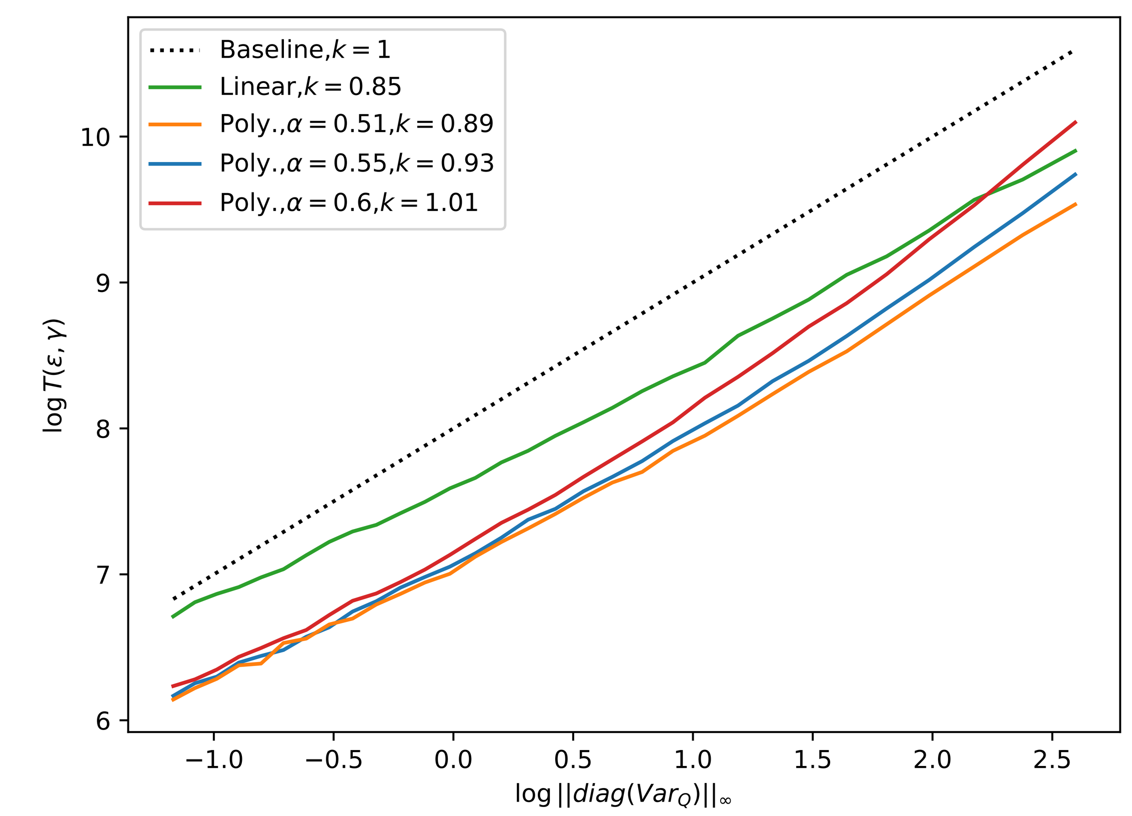

Figure 1:Log-log plots of the sample complexity $T(\varepsilon, \gamma)$ versus the asymptotic variance $\|\mathrm{diag}(\mathsf{Var}_{\boldsymbol{Q}})\|_\infty$ where $T(\varepsilon,\gamma) = \inf\{T : \mathbb{E}\|\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\|_\infty \le \varepsilon\}$ in a γ-discounted MDP.

Conclusion

- Averaged Q-learning achieves both the worst-case and instance-dependent optimality asymptotically.
- The asymptotic variance of averaged Q-learning is optimal among all RAL estimators.
- We established a FCLT that facilitates online statistical inference.

Semiparametric Statistics

- Our MDP model $\mathcal{M}$ has parameter $\theta = (P, R)$.
- The transition P is parametric due to discrete action-state space, while the random reward is totally non-parametric.
- (Unformal) An estimator is regular if its limiting distribution is unaffected by local changes in the data generating process.
- (Unformal) An estimator $\widehat{\boldsymbol{Q}}_T$ is asymptotically linear with a measurable random function $\boldsymbol{\phi}(\boldsymbol{r}_t, \boldsymbol{P}_t) \in \mathbb{R}^D$ such that
$$\sqrt{T}(\widehat{\boldsymbol{Q}}_T - \boldsymbol{Q}^*) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{\phi}(\boldsymbol{r}_t, \boldsymbol{P}_t) + o_{\mathbb{P}}(1), \quad (2)$$
where $\mathcal{D} = \{(\boldsymbol{r}_t, \boldsymbol{P}_t)\}_{t \in [T]}$ is the collected i.i.d. data and $\boldsymbol{\phi}(\cdot,\cdot)$ is referred to as an influence function satisfying that $\mathbb{E}\boldsymbol{\phi}(\boldsymbol{r}_t, \boldsymbol{P}_t) = \boldsymbol{0}$ and $\mathbb{E}\boldsymbol{\phi}(\boldsymbol{r}_t, \boldsymbol{P}_t)\boldsymbol{\phi}(\boldsymbol{r}_t, \boldsymbol{P}_t)^\top$ is finite and nonsingular.

Optimal Variance

We prove the following results.

- RAL = *regular and asymptotically linear.*
- Given the dataset $\mathcal{D}$, the asymptotic variance matrix of any RAL estimator $\widehat{\boldsymbol{Q}}_T$ of $\boldsymbol{Q}^*$ computed from $\mathcal{D}$ satisfying
$$\lim_{T\to\infty} T\mathbb{E}(\widehat{\boldsymbol{Q}}_T - \boldsymbol{Q}^*)(\widehat{\boldsymbol{Q}}_T - \boldsymbol{Q}^*)^\top \succeq \mathsf{Var}_{\boldsymbol{Q}},$$
where $\boldsymbol{A} \succeq \boldsymbol{B}$ means $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite.
- The averaged iterate $\bar{\boldsymbol{Q}}_T$ is the optimal RAL estimator for $\boldsymbol{Q}^*$ due to
$$\sqrt{T}\left(\bar{\boldsymbol{Q}}_T - \boldsymbol{Q}^*\right) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\boldsymbol{I} - \gamma\boldsymbol{P}^{\pi^*})^{-1}\boldsymbol{Z}_t + o_{\mathbb{P}}(1),$$
where $\boldsymbol{Z}_t = (\boldsymbol{r}_t - \boldsymbol{r}) + (\boldsymbol{P}_t - \boldsymbol{P})\boldsymbol{V}^*$.

Averaged Q-Learning iterates has the optimal asymptotic variance matrix.

Functional Central Limit Theorem

We can do statistical inference without estimating $\mathsf{Var}_{\boldsymbol{Q}}$ by using the FCLT.

- Given $\{\boldsymbol{Q}_t\}_{t \in [T]}$, its partial-sum processes is
$$\boldsymbol{\phi}_T(r) := \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tr \rfloor}(\boldsymbol{Q}_t - \boldsymbol{Q}^*), r \in [0,1]. \quad (3)$$
- We assume that (i) $\sup_{s,a}\mathbb{E}R^4(s,a) < \infty$; (ii) π^* is unique; and (iii) $\eta_t = t^{-\alpha}(0.5 < \alpha < 1)$.
- We show the following weak convergence
$$\boldsymbol{\phi}_T(\cdot) \xrightarrow{w} \mathsf{Var}_{\boldsymbol{Q}}^{1/2}\boldsymbol{B}_D(\cdot), \quad (4)$$
where $\mathsf{Var}_{\boldsymbol{Q}}$ is defined in (1) and $\boldsymbol{B}_D$ is the standard D-dim Brownian motion on $[0,1]$.

Online Statistics Inference

- By continuous mapping theorem, for any continuous functional $f : \mathbb{D}[0,1] \to \mathbb{R}$,
$$f(\boldsymbol{\phi}_T) \xrightarrow{d} f(\mathsf{Var}_{\boldsymbol{Q}}^{1/2}\boldsymbol{B}_D).$$
- Once f is scale-invariant, we have $f(\mathsf{Var}_{\boldsymbol{Q}}^{1/2}\boldsymbol{B}_D) = f(\boldsymbol{B}_D)$ and thus
$$\underbrace{f(\boldsymbol{\phi}_T)}_{\substack{\text{(1) a pivotal statistic relying on } \boldsymbol{Q}^* \text{ and } \mathcal{D} \\ \text{(2) can be computed online efficiently}}} \xrightarrow{d} \underbrace{f(\boldsymbol{B}_D)}_{\text{known distribution}}.$$
- Confidence intervals for $\boldsymbol{Q}^*$ can be obtained by inverting some constraint regime on $f(\boldsymbol{\phi}_T)$.
- We find a f following the spirit of t-statistics:
$$f(\boldsymbol{B}) := \boldsymbol{B}(1)^\top\left(\int_0^1 \bar{\boldsymbol{B}}(r)\bar{\boldsymbol{B}}(r)^\top dr\right)^{-1}\boldsymbol{B}(1)$$
with $\bar{\boldsymbol{B}}(r) = \boldsymbol{B}(r) - r\boldsymbol{B}(1)$.
- A numerical illustration: