$Q^\star$ Approximation Schemes for Batch Reinforcement Learning: A Theoretical Comparison

Tengyang Xie, Nan Jiang

University of Illinois at Urbana-Champaign

Value-Function Approximation for Batch RL

- Approximate $Q^\star$ using a restrected function class.
- Given exploratory batch data $\mathcal{D}$ with distribution μ, but no interaction with environment.
- Theoretical foundation of modern reinforcement learning algorithms, e.g., DQN.
- This work: Novel algorithms with better theoretical guarantees.

Theoretical Issues in Value-Function Approximation

(a) Quadratic Dependence on Horizon (i.e., $\mathcal{O}(1/(1-\gamma)^2)$)

Typical performance-to-Bellman-error conversion:

$$J(\pi^\star) - J(\pi_Q) \overset{1/1-\gamma}{\Longrightarrow} \|Q - Q^\star\|_\infty$$

$$\|Q - Q^\star\|_\infty \overset{1/1-\gamma}{\Longrightarrow} \text{algo. objective, e.g., Bellman error}$$

This Paper:

$$J(\pi^\star) - J(\pi_Q) \overset{1/1-\gamma}{\Longrightarrow} Q - \mathcal{T}Q$$

Theoretical Issues in Value-Function Approximation

(b) Characterization of Distribution Shift

Typical "Per-Step" Concentrability Coefficient:

$$C_{\text{per-step}} \coloneqq \sum_{t=0}^{\infty} \beta(t)\, C_t, \quad C_t \coloneqq \max_{\pi} \| w_{d_{\pi,t}/\mu} \|_{\infty},$$

This Paper: stationary-distribution induced concentrability coefficient, e.g.,

$$C_{\infty} \coloneqq \max_{\pi \in \Pi_{\mathcal{Q}}} \| w_{d_{\pi}/\mu} \|_{\infty}.$$

Theoretical Issues in Value-Function Approximation

(c) Function Approximation Assumptions

(Approximate) Closedness under Bellman Update / Low Inherent Bellman Error:

$$\|Q - \mathcal{T}Q\|_{2,\mu} \leq \varepsilon, \quad \forall Q \in \mathcal{Q}$$

This Paper: any (somewhat) weaker alternatives?

Theoretical Issues in Value-Function Approximation

(d) Squared-to-Average Conversion

An Example of Typical Squared-Loss Algorithm (FQI):

$$Q_{k+1} = \operatorname*{argmin}_{Q \in \mathcal{Q}} \mathbb{E}_{\mathcal{D}} \left[\left(Q(s,a) - r - \gamma \max_{a' \in \mathcal{A}} Q_k(s', a') \right)^2 \right]$$

This Paper: Batch RL algorithm with more direct connection to the expected return

Theoretical Issues in Value-Function Approximation

(a) Quadratic Dependence on Horizon (i.e., $\mathcal{O}(1/(1-\gamma)^2)$)

(b) Characterization of Distribution Shift

(c) Function Approximation Assumptions

(d) Squared-to-Average Conversion

Our Contribution: We answer all these questions *positively* by presenting novel analyses of two algorithms, MSBO and MABO.

Telescoping Performance Difference

$$\text{Goal: } J(\pi^\star) - J(\pi_Q) \overset{1/1-\gamma}{\Longrightarrow} Q - \mathcal{T}Q$$

Theorem [Telescoping Performance Difference]: *For any policy π and any $Q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$,*

$$J(\pi) - J(\pi_Q) \leq \frac{\mathbb{E}_{d_\pi}[\mathcal{T}Q - Q]}{1-\gamma} + \frac{\mathbb{E}_{d_{\pi_Q}}[Q - \mathcal{T}Q]}{1-\gamma}.$$

✔ Linear Dependence on Horizon

MSBO — Minimizing *Squared* Bellman Error

Goal: Form an (approximately) unbiased estimate of squared Bellman error $\|Q - \mathcal{T}Q\|_{2,\mu}^2$.

Idea: Capture the over-estimation caused by double sampling.

Minimax Squared Bellman Optimality Error Minimization (MSBO):

$$\widehat{Q} = \operatorname*{argmin}_{Q \in \mathcal{Q}} \max_{f \in \mathcal{F}} \left(\ell_{\mathcal{D}}(Q; Q) - \ell_{\mathcal{D}}(f; Q) \right),$$

$$\ell_{\mathcal{D}}(f; Q) \coloneqq \mathbb{E}_{\mathcal{D}} \left[\left(f(s, a) - r - \gamma \max_{a' \in \mathcal{A}} Q(s', a') \right)^2 \right].$$

MSBO — Minimizing *Squared* Bellman Error

(Improved) Performance Bound of MSBO: Let $\widehat{Q}$ be the output of MSBO. W.p. at least $1-\delta$,

$$\max_{\pi \in \Pi_{\mathcal{Q}}} J(\pi) - J(\pi_{\widehat{Q}}) \le \frac{2\sqrt{2C_{\text{eff}}}}{1-\gamma}\left(\sqrt{\varepsilon_{\mathcal{Q}}^{\text{sq}}} + \sqrt{\varepsilon_{\mathcal{Q},\mathcal{F}}^{\text{sq}}}\right) +$$

$$\frac{\sqrt{C_{\text{eff}}}}{1-\gamma}\mathcal{O}\left(\sqrt{\frac{V_{\max}^2 \ln\frac{|\mathcal{Q}||\mathcal{F}|}{\delta}}{n}} + \sqrt[4]{\frac{V_{\max}^2 \ln\frac{|\mathcal{Q}|}{\delta}}{n}\varepsilon_{\mathcal{Q}}^{\text{sq}}} + \sqrt[4]{\frac{V_{\max}^2 \ln\frac{|\mathcal{Q}||\mathcal{F}|}{\delta}}{n}\varepsilon_{\mathcal{Q},\mathcal{F}}^{\text{sq}}}\right),$$

where $C_{\text{eff}} \coloneqq \max_{\pi \in \Pi_{\mathcal{Q}}} \|{}^{d_\pi}/_{\mu}\|_{2,\mu}^2$, $\varepsilon_{\mathcal{Q}}^{\text{sq}} \coloneqq \min_{Q \in \mathcal{Q}} \|Q - \mathcal{T}Q\|_{2,\mu}^2$, and $\varepsilon_{\mathcal{Q},\mathcal{F}}^{\text{sq}} \coloneqq \max_{Q \in \mathcal{Q}} \min_{f \in \mathcal{F}} \|f - \mathcal{T}Q\|_{2,\mu}^2$.

- ✔ Linear Dependence on Horizon
- ✔ Tight and Elegant Characterization of Distribution Shift
- ✘ Function Approximation Assumptions
- ✘ Squared-to-Average Conversion

MABO — Minimizing *Average* Bellman Error

Goal: Approximate $Q^\star$ by directly estimating the average Bellman error.

Idea: Minimizing average Bellman error $|\mathbb{E}_\pi[Q - \mathcal{T}Q]|$ with any policy π (thanks to the performance difference telescoping).

Minimax Average Bellman Optimality Error Minimization (MABO):

$$\widehat{Q} = \operatorname*{argmin}_{Q \in \mathcal{Q}} \max_{w \in \mathcal{W}} \quad |\mathcal{L}_{\mathcal{D}}(Q, w)|,$$

$$\mathcal{L}_{\mathcal{D}}(Q, w) \coloneqq \mathbb{E}_{\mathcal{D}}\left[w(s, a)\left(r + \gamma \max_{a'} Q(s', a') - Q(s, a)\right)\right].$$

MABO — Minimizing *Average* Bellman Error

Performance Bound of MABO: Let $\widehat{Q}$ be the output of MABO. W.p. $1-\delta$,

$$\max_{\pi \in \Pi_{\mathcal{Q}}} J(\pi) - J(\pi_{\widehat{Q}}) \leq \frac{2}{1-\gamma}\left(\varepsilon_{\mathcal{Q}}^{\text{avg}} + \varepsilon_{\mathcal{Q},\mathcal{W}}^{\text{avg}} + \varepsilon_{\text{stat},n}\right),$$

where $\varepsilon_{\mathcal{Q}}^{\text{avg}} \coloneqq \min_{Q \in \mathcal{Q}} \max_{w \in \mathcal{W}} |\mathbb{E}_{\mu}[w \cdot (\mathcal{T}Q - Q)]|$,

$\varepsilon_{\mathcal{Q},\mathcal{W}}^{\text{avg}} \coloneqq \max_{\pi \in \Pi_{\mathcal{Q}}} \inf_{w \in \text{sp}(\mathcal{W})} \max_{Q \in \mathcal{Q}} |\mathbb{E}_{\mu}[(w_{d_{\pi}/\mu} - w) \cdot (\mathcal{T}Q - Q)]|$,

$$\varepsilon_{\text{stat},n} \coloneqq 2V_{\max}\sqrt{\frac{2C_{\text{eff},\mathcal{W}} \ln \frac{2|\mathcal{Q}||\mathcal{W}|}{\delta}}{n}} + \frac{4C_{\infty,\mathcal{W}} V_{\max} \ln \frac{2|\mathcal{Q}||\mathcal{W}|}{\delta}}{3n},$$

$C_{\text{eff},\mathcal{W}} \coloneqq \max_{w \in \mathcal{W}} \|w\|_{2,\mu}^2, \quad C_{\infty,\mathcal{W}} \coloneqq \max_{w \in \mathcal{W}} \|w\|_{\infty}.$

- ✔ Linear Dependence on Horizon
- ✔ Tight and Elegant Characterization of Distribution Shift
- ✔ Function Approximation Assumptions
- ✔ Squared-to-Average Conversion

Thanks!