RL with continuous action spaces

General function approximation continued…

Aarti Singh

Machine Learning 10-734
Nov 11, 2025

Slides courtesy: Wen Sun

Average Bellman error – Q version

Weaker notion of Bellman error:

Evaluate g-approximation of Q using a policy π_f

$$\mathscr{E}(g; f, h) = \mathbb{E}_{s_h, a_h \sim d_h^{\pi_f}} \left[g(s_h, a_h) - r(s_h, a_h) - \mathbb{E}_{s_{h+1} \sim P(\cdot|s_h, a_h)} \left[\max_{a \in \mathcal{A}} g(s_{h+1}, a) \right] \right]$$

f: defines roll-in distribution over s_h, a_h

We know that $\mathscr{E}(Q^\star; f, h) = 0, \forall f$

Hence, any g such that $\mathscr{E}(g; f, h) \neq 0$, is an incorrect $Q^\star$ approximator

Average Bellman error – V version

Evaluate average Bellman wrt *V* function induced by g as well:

$$\mathscr{E}(g;f,h) = \mathbb{E}_{s_h \sim d_h^{\pi_f}} \left[V_g(s_h) - r(s_h, \pi_g(s_h)) - \mathbb{E}_{s_{h+1} \sim P(\cdot|s_h, \pi_g(s_h))} \left[V_g(s_{h+1}) \right] \right]$$

f: defines roll-in distribution over s_h, a_h

and $V_g(s) = max_a\, g(s,a)$.

Again we have $\mathscr{E}(Q^\star;f,h) = 0, \forall f$

(because: $V_{Q^\star}(s) - r(s, \pi_{Q^\star}(s)) - \mathbb{E}_{s' \sim P_h(.|s,\pi_{Q^\star}(s))} V_{Q^\star}(s') = 0$)

Hence, any g such that $\mathscr{E}(g;f,h) \neq 0$, is an incorrect $Q^\star$ approximator

Estimating Q-Bellman error (generative setting)

Recall our hypothesis class $\mathcal{F}$, where each $g \in \mathcal{F}$ is in the form of $g(s, a)$

For Q-Bellman rank, we define Bellman error loss as:

$$\ell(s_h, a_h, s'_{h+1}, g) = g(s_h, a_h) - r(s_h, a_h) - \max_{a'} g(s_{h+1}, a')$$

If we had a dataset $\mathcal{D} := \{s_h, a_h, s_{h+1}\}$ where $s_h, a_h \sim d_h^{\pi_f}, s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$

then $\forall g : \mathbb{E}_{\mathcal{D}}[\ell(s_h, a_h, s_{h+1}, g)]$ is an unbiased est of $\mathcal{E}(g; f, h)$

Estimating V-Bellman error (generative setting)

Recall our hypothesis class $\mathcal{F}$, where each $g \in \mathcal{F}$ is in the form of $g(s, a)$

For V-Bellman rank, we define Bellman error loss as:

$$\ell(s_h, a_h, s'_{h+1}, g) = \frac{\mathbf{1}\{a_h = \pi_g(s_h)\}}{1/A} \left(g(s_h, a_h) - r(s_h, a_h) - \max_{a'} g(s_{h+1}, a') \right)$$

If we had a dataset $\mathcal{D} := \{s_h, a_h, s_{h+1}\}$ where $s_h \sim d_h^{\pi_f}, a_h \sim U(\mathcal{A}),$
$s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$

then $\forall g : \mathbb{E}_{\mathcal{D}}[\ell(s_h, a_h, s_{h+1}, g)]$ is an unbiased est of $\mathcal{E}(g; f, h)$

Estimating Bellman rank

Uniform convergence style assumption on our hypothesis class $\mathcal{F}$:

Given any distribution $\nu \in \Delta(S \times A \times S)$, and m i.i.d samples $\{s_i, a_i, s_i'\}$ from ν, w/ probability at least $1 - \delta$,

$$\forall g : \left| \mathbb{E}_{\nu} \ell(s, a, s', g) - \mathbb{E}_{\mathcal{D}} \ell(s, a, s', g) \right| \leq \varepsilon_{gen}(m, \mathcal{F}, \delta)$$

Example: when $\mathcal{F}$ is discrete (for B-rank loss), Hoeffding + union bound over $\mathcal{F}$ implies:

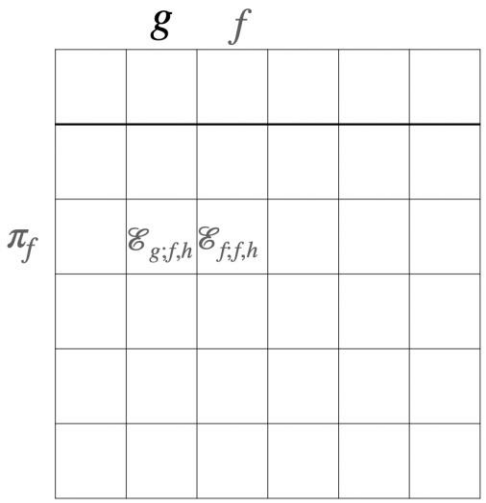
$$\varepsilon_{gen}(m, \mathcal{F}, \delta) := 2H\sqrt{\frac{\ln(|\mathcal{F}|/\delta)}{m}}$$

Bellman rank

$\exists$ two mappings $W_h : \mathcal{F} \mapsto \mathbb{R}^d, \quad X_h : \mathcal{F} \mapsto \mathbb{R}^d$ (d = Bellman-rank)

s.t. $\forall f, g \in \mathcal{F} : \mathcal{E}(g; f, h) = \langle W_h(g), X_h(f) \rangle$

$\forall h : \mathcal{E}_h \in \mathbb{R}^{|\mathcal{F}| \times |\mathcal{F}|}$



Rank of this matrix = Bellman rank

Note: we just assume the existence of W, X, but they are unknown

Algorithm under Bellman rank

- Works for general function approximation with low Bellman rank
- Gives optimism without adding (nonlinear) bonus

Bilinear-UCB

A general algorithm under Bellman rank that can learn an ϵ near optimal policy with number of samples

$$\text{e.g., poly}(H, \text{b-rank}, \ln(|\mathcal{F}|), 1/\epsilon^2)$$

Bilinear-UCB

At iteration t :

Select $f_t = \arg\max_{g \in \mathcal{F}} V_g(s_0)$

$$\text{s.t., } \forall h : \sum_{i=0}^{t-1} \left(\mathbb{E}_{\mathcal{D}_{h,i}} [\ell(s_h, a_h, s_{h+1}, g)] \right)^2 \leq R^2$$

Bilinear-UCB

At iteration t :

Select $f_t = \arg\max_{g \in \mathcal{F}} V_g(s_0)$

$$\text{s.t., } \forall h : \sum_{i=0}^{t-1} \left(\mathbb{E}_{\mathcal{D}_{h,i}} [\ell(s_h, a_h, s_{h+1}, g)] \right)^2 \leq R^2$$

For all h, create $\mathcal{D}_{h,t} = \{s_h, a_h, s_{h+1}\}$ w/ m triples, where:

- For Q-B rank case: $s_h, a_h \sim d_h^{\pi_{f_t}}, s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$
- For V-B rank case: $s_h \sim d_h^{\pi_{f_t}}, a_h \sim U(A), s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$

Bilinear-UCB

$$\text{Select } f_t = \arg\max_{g \in \mathcal{F}} V_g(s_0) \quad \text{s.t., } \forall h : \sum_{i=0}^{t-1} \left(\mathbb{E}_{\mathcal{D}_{h,i}}[\ell(s_h, a_h, s_{h+1}, g)] \right)^2 \leq R^2$$

1. When the batch size ($|\mathcal{D}_{h,i}|$) is large,

$$\mathbb{E}_{\mathcal{D}_{h,i}} \ell(s_h, a_h, s_{h+1}, g) \to \mathcal{E}(g; f_i, h)$$

2. We know that $\sum_{i=1}^{t-1} \mathcal{E}(f^\star; f_i, h) = 0$

3. By properly setting batch size and R, we eliminate wrong hypothesis, but keep $f^\star$

4. This gives optimism: $V_{f_t}(s_0) \geq V_{f^\star}(s_0) := V^\star(s_0)$

Optimism allows explore and exploit tradeoff!

Analysis of Bilinear-UCB

Step 1: proving optimism via showing $f^\star$ is always a feasible solution (whp)

Recall constraint: $\forall h : \sum_{i=0}^{t-1} \left(\mathbb{E}_{\mathcal{D}_{h,i}}[\ell(s_h, a_h, s_{h+1}, g)] \right)^2 \leq R^2$

Lemma: set $R = \sqrt{T} \cdot \varepsilon_{gen}(m, \mathcal{F}, \delta/TH)$,

W/ prob $1 - \delta$, we have $f^\star$ being a feasible solution for all the T iterations;

Proof: Consider any iteration $i < t$:

$$|\mathbb{E}_{\mathcal{D}_{i,h}} \ell(s_h, a_h, s_{h+1}, f^\star) - \mathcal{E}(f^\star; f_i, h)| \leq \varepsilon_{gen}$$

$$(\mathbb{E}_{\mathcal{D}_{i,h}} \ell(s_h, a_h, s_{h+1}, f^\star))^2 \leq \varepsilon_{gen}^2 \qquad \text{since } \mathcal{E}(f^\star; f_i, h) = 0$$

$$\sum_{i=0}^{t-1} \left(\mathbb{E}_{\mathcal{D}_{i,h}} \ell(s_h, a_h, s_{h+1}, f^\star) \right)^2 \leq t\varepsilon_{gen}^2 \leq T\varepsilon_{gen}^2 := R^2$$

Analysis of Bilinear-UCB

Step 1: proving optimism via showing $f^\star$ is always a feasible solution (whp)

The fact that $f^\star$ being feasible $\Rightarrow$ optimism, i.e., $\forall t, V_{f_t}(s_0) \geq V_{f^\star}(s_0) := V^\star(s_0)$

Proof:

Recall the objective function:

$$\text{Select } f_t = \arg\max_{g \in \mathcal{F}} V_g(s_0) \quad \text{s.t., } \forall h : \sum_{i=0}^{t-1} \left(\mathbb{E}_{\mathcal{D}_{h,i}}[\ell(s_h, a_h, s_{h+1}, g)] \right)^2 \leq R^2$$

Analysis of Bilinear-UCB

Step 2: Using optimism to upper bound per-episode regret:

$$\text{Optimism} \Rightarrow V^{\star}(s_0) - V^{\pi_{f_t}}(s_0) \leq V_{f_t}(s_0) - V^{\pi_{f_t}}(s_0)$$

Lemma:

$$V_{f_t}(s_0) - V^{\pi_{f_t}}(s_0) = \sum_{h=0}^{H-1} \mathbb{E}_{s_h, a_h \sim d_h^{\pi_{f_t}}} \left[f_t(s_h, a_h) - r(s_h, a_h) - \mathbb{E}_{s_{h+1} \sim P_h(s_h, a_h)} \max_{a'} f_t(s_{h+1}, a') \right]$$

$$= \sum_{h=0}^{H-1} \mathscr{E}(f_t; f_t, h) = \sum_{h=0}^{H-1} W_h(f_t)^{\top} X_h(f_t)$$

Proof:

$$V_{f_t}(s_0) - V^{\pi_{f_t}}(s_0) = \mathbb{E}_{a_0 \sim \pi_{f_t}(s_0)}[f_t(s_0, a_0)] - \mathbb{E}_{a_h \sim \pi_{f_t}(s_h),\, s_{h+1} \sim P(.|s_h, a_h)}[\textstyle\sum_{h=0}^{H-1} r(s_h, a_h)]$$

$$= \mathbb{E}_{a_h \sim \pi_{f_t}(s_h),\, s_{h+1} \sim P(.|s_h, a_h)}[\textstyle\sum_{h=0}^{H-1} (f_t(s_h, a_h) - r(s_h, a_h) - f_t(s_{h+1}, a_{h+1}))]$$

telescoping sum since $f_t(s_H, a_H) = 0$

$$= \textstyle\sum_{h=0}^{H-1} \mathbb{E}_{a_h, s_h \sim d^{\pi_{f_t}}} [f_t(s_h, a_h) - r(s_h, a_h) - \mathbb{E}_{s_{h+1} \sim P(.|s_h, a_h)}[\max_{a'} f_t(s_{h+1}, a')]]$$

since $a_{h+1} = \arg\max_{a'} f_t(s_{h+1}, a')$

Analysis of Bilinear-UCB

Step 2: Using optimism to upper bound per-episode regret:

$$V^{\star}(s_0) - V^{\pi_{f_t}}(s_0) = \sum_{h=0}^{H-1} \mathbb{E}_{s_h, a_h \sim d_h^{\pi_{f_t}}} \left[f_t(s_h, a_h) - r(s_h, a_h) - \mathbb{E}_{s_{h+1} \sim P_h(s_h, a_h)} \max_{a'} f_t(s_{h+1}, a') \right]$$

$$= \sum_{h=0}^{H-1} \mathscr{E}(f_t; f_t, h) = \sum_{h=0}^{H-1} W_h(f_t)^{\top} X_h(f_t)$$

Define ``feature" covariance matrix $\Sigma_{t,h} = \sum_{i=0}^{t-1} X_h(f_i) X_h(f_i)^{\top} + \lambda I$

Cauchy-Schwartz implies $V^{\star}(s_0) - V^{\pi_{f_t}}(s_0) \leq \sum_{h=0}^{H-1} \left\| W_h(f_t) \right\|_{\Sigma_{t,h}} \| X_h(f_t) \|_{\Sigma_{t,h}^{-1}}$

Analysis of Bilinear-UCB

Summary so far, after optimism + per-episode regret decomposition, we get:

Define ``feature" covariance matrix $\Sigma_{t,h} = \sum_{i=0}^{t-1} X_h(f_i) X_h(f_i)^\top + \lambda I$

$$\forall t: \; V^\star(s_0) - V^{\pi_{f_t}}(s_0) \leq \sum_{h=0}^{H-1} \left\| W_h(f_t) \right\|_{\Sigma_{t,h}} \|X_h(f_t)\|_{\Sigma_{t,h}^{-1}}$$

Similar to linUCB, using elliptical potential lemma:

$$\|X_h(f_t)\|^2_{\Sigma_{t,h}^{-1}} \leq \exp\left(\frac{Hd}{T} \ln\left(1 + \frac{TB_X^2}{d\lambda}\right)\right) - 1,$$

where $\|X_h(f)\|_2 \leq B_X \quad \forall f \in \mathcal{F}$

Similarly, let $\|W_h(f)\|_2 \leq B_W \quad \forall f \in \mathcal{F}$

Analysis of Bilinear-UCB

Summary so far, after optimism + per-episode regret decomposition, we get:

Define ``feature" covariance matrix $\Sigma_{t,h} = \sum_{i=0}^{t-1} X_h(f_i) X_h(f_i)^\top + \lambda I$

$$\forall t: \; V^\star(s_0) - V^{\pi_{f_t}}(s_0) \leq \sum_{h=0}^{H-1} \left\| W_h(f_t) \right\|_{\Sigma_{t,h}} \|X_h(f_t)\|_{\Sigma_{t,h}^{-1}}$$

$$\forall h: \sum_{i=0}^{t-1} \left(\mathbb{E}_{\mathscr{D}_{h,i}}[\ell(s_h, a_h, s_{h+1}, f_t)] \right)^2 \leq R^2$$

$$\Rightarrow \forall h: \sum_{i=0}^{t-1} \mathscr{E}(f_t; f_i, h)^2 \leq \sum_{i=0}^{t-1} 2 \left(\mathscr{E}(f_t; f_i, h) - \mathbb{E}_{\mathscr{D}_{i,h}} \ell(s_h, a_h, s_{h+1}, f_t) \right)^2 + \sum_{i=0}^{t-1} 2 \left(\mathbb{E}_{\mathscr{D}_{i,h}} \ell(s_h, a_h, s_{h+1}, f_t) \right)^2$$

$$\Rightarrow \forall h: \sum_{i=0}^{t-1} \mathscr{E}(f_t; f_i, h)^2 \leq 4T\varepsilon_{gen}^2 \quad \Rightarrow \forall h: \sum^{t-1} \left(W_h(f_t)^\top X_h(f_i) \right)^2 \leq 4T\varepsilon_{gen}^2$$

$$\Rightarrow \forall h: \left\| W_h(f_t) \right\|_{\Sigma_{t,h}}^2 \leq 4T\varepsilon_{gen}^2 + \lambda B_W^2$$

Analysis of Bilinear-UCB

$$\forall t: \; V^{\star}(s_0) - V^{\pi_{f_t}}(s_0) \leq \sum_{h=0}^{H-1} \left\| W_h(f_t) \right\|_{\Sigma_{t,h}} \|X_h(f_t)\|_{\Sigma_{t,h}^{-1}}$$

$$\leq H\sqrt{4\lambda B_W^2 + 4T\varepsilon_{gen}^2}\sqrt{\exp\left(\frac{Hd}{T}\ln(1+\frac{TB_X^2}{d\lambda})\right) - 1}$$

Set $1/\lambda = B_W^2/\varepsilon_{gen}^2 + 1/B_X^2$ And $T = \left\lceil 2Hd\ln\left(4Hd\left(\frac{B_X^2 B_W^2}{\varepsilon_{gen}^2}+1\right)\right)\right\rceil$

$$\leq H\sqrt{2\left(4\lambda B_W^2 + 4T\varepsilon_{gen}^2\right)}$$

$$\leq 5\varepsilon_{gen}\sqrt{dH^3\ln\left(4Hd\left(\frac{B_X^2 B_W^2}{\varepsilon_{gen}^2}+1\right)\right)}.$$

Analysis of Bilinear-UCB

$$\forall t: \; V^{\star}(s_0) - V^{\pi_{f_t}}(s_0) \leq \sum_{h=0}^{H-1} \left\| W_h(f_t) \right\|_{\Sigma_{t,h}} \|X_h(f_t)\|_{\Sigma_{t,h}^{-1}}$$

Regret bound

$$\leq 5\varepsilon_{gen} \sqrt{dH^3 \ln\left(4Hd\left(\frac{B_X^2 B_W^2}{\varepsilon_{gen}^2} + 1\right)\right)}.$$

Example: when $\mathcal{F}$ is discrete (for B-rank loss), Hoeffding + union bound over $\mathcal{F}$ implies:

$$\varepsilon_{gen}(m, \mathcal{F}, \delta) := 2H\sqrt{\frac{\ln(|\mathcal{F}|/\delta)}{m}}$$

eps-optimal with sample complexity , m ~ poly (H, d, ln(| $\mathcal{F}$ |),1/eps^2)

General function approximation

Summary:

- General non-linear function approximation
- Hardness - Exponential complexity with continuous arms
- Poly sample complexity under Bellman rank, low-rank MDP, Bellman completeness
- Bilinear-UCB algorithm

Towards model-free RL

Open questions:

- Computational complexity
- Online setting (needs additional assumptions)

Instead of parametrizing model/value functions, can we parametrize policies directly?