Generalization in Large scale MDPs

Sham Kakade and Wen Sun

CS 6789: Foundations of Reinforcement Learning

Recap: Bellman error of Q

We define **average** Bellman error of a Q-estimate g below:

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

Recap: Bellman error of Q

We define **average** Bellman error of a Q-estimate g below:

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

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

Recap: Bellman error of the associated V functions

We can define **average** Bellman error wrt the V-function induced by g as well:

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

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

$$\pi_g(s) = \arg\max_a g(s, a)$$

Recap: Bellman error of the associated V functions

We can define **average** Bellman error wrt the V-function induced by g as well:

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

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

Recap: Bellman error of the associated V functions

We can define **average** Bellman error wrt the V-function induced by g as well:

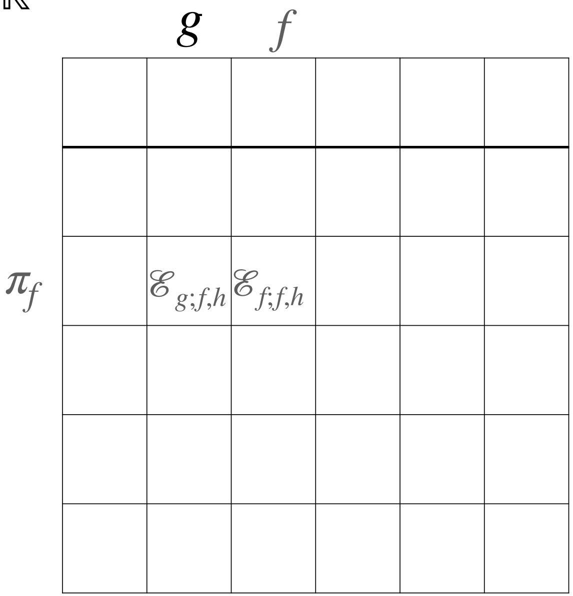
$$\mathscr{E}(g; f, h) = \mathbb{E}_{s_h \sim d_h^{\pi_f}} \left[V_g(s_h) - r(s_h, \pi_g(s_h)) - \mathbb{E}_{s_{h+1} \sim P(\cdot|s_h, \pi_g(s_h))} \left[V_g(s_{h+1}) \right] \right]$$

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

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

Recap: The Q / V-Bellman rank

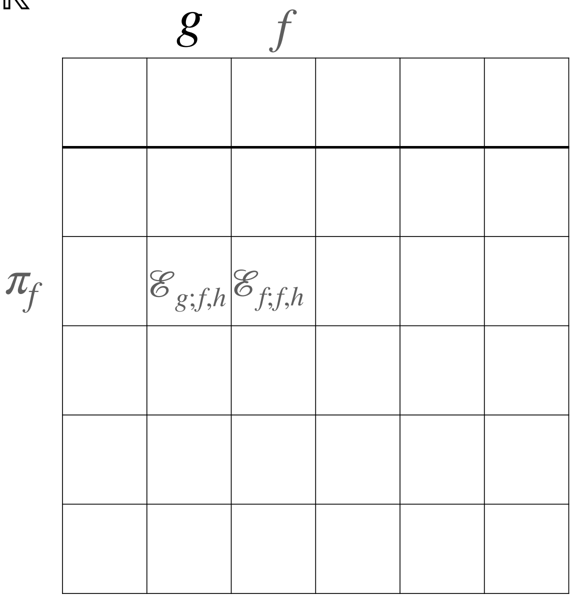
$\forall h : \mathscr{E}_h \in \mathbb{R}^{|\mathscr{F}| \times |\mathscr{F}|}$



Rank of this Matrix is defined as Bellman Rank

Recap: The Q / V-Bellman rank

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



There are two mappings

$W_h : \mathscr{F} \mapsto \mathbb{R}^d, \quad X_h : \mathscr{F} \mapsto \mathbb{R}^d$

(d = Q/V Bellman-rank)

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

Rank of this Matrix is defined as Bellman Rank

Recap: Many examples have low Bellman rank

1. Linear Bellman completion (including linear and tabular MDPs, and LQR)

2. Linear $Q^\star$ & $V^\star$ (captures the $Q^\star$-state abstraction)

3. Low-rank MDPs (unknown representation that needs to be learned)

4. Many others: Reactive POMDPs, Contextual bandit, Low-occupancy measures…

Question for Today

Can we design a universal algorithm that learns efficiently for MDPs w/ low-Q/V Bellman rank?

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

Outline for Today

1. The Bilinear-UCB algorithm (BLin-UCB)

2. Theoretical Guarantee and analysis of BLin-UCB

For Q-Bellman rank case:

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

For Q-Bellman rank case:

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

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

For Q-Bellman rank case:

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

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

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

For Q-Bellman rank case:

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

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

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

Test policy

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

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

$$:= \frac{1}{|D|} \sum_{i=1}^{|D|} \ell(s^i, a^i, s'^i, g)$$

$$\mathbb{E}_{s,a \sim d_h^{\pi_f}} g(s,a) - r - \max_{a'} g(s', a')$$

For V-Bellman rank case:

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

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

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

For V-Bellman rank case:

Recall our hypothesis class $\mathscr{F}$, where each $g \in \mathscr{F}$ 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 $\mathscr{D} := \{s_h, a_h, s_{h+1}\}$ where $s_h \sim d_h^{\pi_f}, a_h \sim U(\mathscr{A}), s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$

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

The Algorithm:

At iteration t :

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

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

The Algorithm:

At iteration t :

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

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

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

The Algorithm:

At iteration t :

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

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: $s_h, a_h \sim d_h^{\pi_{f_t}}, s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$

The Algorithm:

$D_{h,i} \leftarrow \{s_h, a_h, s_{h+1}\}$ from π_{f_i}

At iteration t :

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

$\Rightarrow \mathcal{E}(g; f_i, h)$

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: $s_h \sim d_h^{\pi_{f_t}}, a_h \sim U(A), s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$

Intuition behind the algorithm:

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

Intuition behind the algorithm:

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

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

Intuition behind the algorithm:

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

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

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

Intuition behind the algorithm:

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

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 $f^\star$

Intuition behind the algorithm:

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

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: $V_{f_t}(s_0) \ge V_{f^\star}(s_0) := V^\star(s_0)$

Intuition behind the algorithm:

Select $g_t = \arg\max_{g \in \mathcal{F}} V_g(s_0)$ 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: $V_{f_t}(s_0) \geq V_{f^\star}(s_0) := V^\star(s_0)$

3. Optimism allows explore and exploit tradeoff!

Outline for Today

✓ 1. The Bilinear-UCB algorithm (BLin-UCB)

2. Theoretical Guarantee and analysis of BLin-UCB

Analysis of BLin-UCB

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

Analysis of BLin-UCB

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

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 $1 - \delta$,

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

$\to 0$, as $m \to \infty$

Analysis of BLin-UCB

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

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 $1 - \delta$,

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

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

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

$\mathscr{F}$ is linear

$\varepsilon_{gen} \approx \frac{d}{\sqrt{m}}$

Analysis of BLin-UCB

Bellman-Rank

After running BLin-UCB for $T = \widetilde{O}\left(Hd\right)$ many iterations, there exists a policy among T many policies, such that:

$$V^\star(s_0) - V^\pi(s_0) \leq \widetilde{O}\left(\varepsilon_{gen}\left(m, \mathscr{F}, \delta/(TH)\right) \cdot \sqrt{dH^3}\right) \leq \varepsilon$$

(# of trajectories used: mHT)

When F is discrete:

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

$$H\sqrt{\frac{\ln(|F|/\delta)}{m}}\sqrt{dH^3} \leq \varepsilon$$

Analysis of BLin-UCB

Example: discrete (but large) hypothesis class $\mathscr{F}$ for Q-Bellman rank

W/ prob $1-\delta$, BLin-UCB learns a policy with $V^\star - V^\pi \leq \epsilon$, w/ # of trajectories:

$$\widetilde{O}\left(\frac{H^6 d^2 \ln\left(|\mathscr{F}|/\delta\right)}{\epsilon^2}\right)$$

Analysis of BLin-UCB

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

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

$\approx \mathcal{E}(g; f_i; h)$

Analysis of BLin-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;

Analysis of BLin-UCB

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

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

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

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

Consider any iteration $i < t$:

Analysis of BLin-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;

$f^\star := Q^\star$

Consider any iteration $i < t$:

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

Analysis of BLin-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;

Consider any iteration $i < t$:

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

$$(\mathbb{E}_{\mathcal{D}_{i,h}} \ell(s_h, a_h, s_{h+1}, f^\star))^2 \leq 2 \left(\mathbb{E}_{\mathcal{D}_{i,h}} \ell(s_h, a_h, s_{h+1}, f^\star) - \mathcal{E}(f^\star; f_i, h) \right)^2 + 2(\mathcal{E}(f^\star; f_i, h))^2$$

$(a+b)^2 \leq 2a^2 + 2b^2$ ✓

$\leq \varepsilon_{gen}^2$

$= 0$

Analysis of BLin-UCB

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

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

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

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

Consider any iteration $i < t$:

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

$$(\mathbb{E}_{\mathscr{D}_{i,h}} \ell(s_h, a_h, s_{h+1}, f^\star))^2 \leq 2 \left(\mathbb{E}_{\mathscr{D}_{i,h}} \ell(s_h, a_h, s_{h+1}, f^\star) - \mathscr{E}(f^\star; f_i, h) \right)^2 + 2(\mathscr{E}(f^\star; f_i, h))^2 \leq 2\varepsilon_{gen}^2$$

Analysis of BLin-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;

Consider any iteration $i < t$:

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

$$(\mathbb{E}_{\mathcal{D}_{i,h}} \ell(s_h, a_h, s_{h+1}, f^\star))^2 \leq 2 \left(\mathbb{E}_{\mathcal{D}_{i,h}} \ell(s_h, a_h, s_{h+1}, f^\star) - \mathcal{E}(f^\star; f_i, h) \right)^2 + 2(\mathcal{E}(f^\star; f_i, h))^2 \leq 2\varepsilon_{gen}^2$$

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

Analysis of BLin-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., $\forall t, V_{f_t}(s_0) \geq V_{f^\star}(s_0) := V^\star(s_0)$

Analysis of BLin-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 BLin-UCB

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

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

Lemma:

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

$\mathcal{E}(f_t; f_t, h)$

Analysis of BLin-UCB

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

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

Lemma:

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

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

Analysis of BLin-UCB

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

Lemma:

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

Analysis of BLin-UCB

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

Lemma:

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

Key trick: telescoping

Analysis of BLin-UCB

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

Lemma:

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

Key trick: telescoping

$$h = 0 : \quad f_t(s_0, a_0) - \mathbb{E}_{s_1 \sim d_1^{\pi_{f_t}}} \max_{a'} f_t(s_1, a')$$

Analysis of BLin-UCB

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

Lemma:

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

Key trick: telescoping

$$h = 0: \quad f_t(s_0, a_0) - \mathbb{E}_{s_1 \sim d_1^{\pi_{f_t}}} \max_{a'} f_t(s_1, a')$$

$a' = \pi_{f_t}(s_1)$

$$h = 1: \quad \mathbb{E}_{s_1 \sim d_1^{\pi_{f_t}}, a_1 = \pi_{f_t}(s_1)} f_t(s_1, a_1) - \mathbb{E}_{s_2 \sim d_2^{\pi_{f_t}}} \max_{a'} f_t(s_2, a')$$

Analysis of BLin-UCB

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

Lemma:

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

Key trick: telescoping

$$h = 0 : \quad f_t(s_0, a_0) - \mathbb{E}_{s_1 \sim d_1^{\pi_{f_t}}} \max_{a'} f_t(s_1, a')$$

$$h = 1 : \quad \mathbb{E}_{s_1 \sim d_1^{\pi_{f_t}}, a_1 = \pi_{f_t}(s_1)} f_t(s_1, a_1) - \mathbb{E}_{s_2 \sim d_2^{\pi_{f_t}}} \max_{a'} f_t(s_2, a')$$

$$h = 2, \ldots$$

Analysis of BLin-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 $\Sigma_{t,h} = \sum_{i=0}^{t-1} X_h(f_i) X_h(f_i)^\top + \lambda I$

Via CS inequality:

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

Analysis of BLin-UCB

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

Define ``feature'' covariance matrix $\Sigma_{t,h} = \sum_{i=0}^{t-1} X_h(f_i) X_h(f_i)^\top + \lambda I$

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

V^{π} π_{f_t}

$f : \; V_f := \max_a f(\cdot, a)$

Analysis of BLin-UCB

Step 3: argue that we make progress whenever π_{f_t} is not good…

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

Analysis of BLin-UCB

Step 3: argue that we make progress whenever π_{f_t} is not good…

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

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

$\approx \mathcal{E}(g; f_i; h)$

Analysis of BLin-UCB

Step 3: argue that we make progress whenever π_{f_t} is not good…

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

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

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

Analysis of BLin-UCB

Step 3: argue that we make progress whenever π_{f_t} is not good…

$$V^\star(s_0) - V^{\pi_{f_t}}(s_0) \le \sum_{h=0}^{H-1} \| W_h(f_t) \|_{\Sigma_{t,h}} \|X_h(f_t)\|_{\Sigma_{t,h}^{-1}}$$

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

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

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

$\le 2\varepsilon_{gen}^2$

$(a+b)^2 \le 2a^2 + 2b^2$

$\le 2T\varepsilon_{gen}$

Analysis of BLin-UCB

Step 3: argue that we make progress whenever π_{f_t} is not good…

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

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

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

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

$$\Rightarrow \forall h : \sum_{i=0}^{t-1} \mathcal{E}(f_t; f_i, h)^2 \le 4T\varepsilon_{gen}^2$$

Analysis of BLin-UCB

Step 3: argue that we make progress whenever π_{f_t} is not good…

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

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

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

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

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

Analysis of BLin-UCB

$\max_{h,f} \| W_h(f) \|_2 \le B_W$

Step 3: argue that we make progress whenever π_{f_t} is not good…

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

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

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

$W_h(f_t)^\top \sum_{i=0}^{t-1} X_h(f_i) X_h(f_i)^\top W_h(f_t)$

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

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

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

Analysis of BLin-UCB

Step 3: argue that we make progress whenever π_{f_t} is not good…

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

$$\le \sum_{h=0}^{H-1} \sqrt{4T\varepsilon_{gen}^2 + \lambda B_W^2} \left\| X_h(f_t) \right\|_{\Sigma_{t,h}^{-1}}$$

Analysis of BLin-UCB

Step 3: argue that we make progress whenever π_{f_t} is not good…

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

$$\leq \sum_{h=0}^{H-1} \sqrt{4T\varepsilon_{gen}^2 + \lambda B_W^2} \left\| X_h(f_t) \right\|_{\Sigma_{t,h}^{-1}}$$

If $V^\star(s_0) - V^{\pi_{f_t}}(s_0) \geq \epsilon$,

Analysis of BLin-UCB

Step 3: argue that we make progress whenever π_{f_t} is not good…

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

$$\leq \sum_{h=0}^{H-1} \sqrt{4T\varepsilon_{gen}^2 + \lambda B_W^2} \| X_h(f_t) \|_{\Sigma_{t,h}^{-1}}$$

If $V^\star(s_0) - V^{\pi_{f_t}}(s_0) \geq \epsilon$,

Then, we know that $\exists h$, such that $\|X_h(f_t)\|_{\Sigma_{t,h}^{-1}} \geq \epsilon / \left(\sqrt{4T\varepsilon_{gen}^2 + \lambda B_W^2} \cdot H\right)$

Analysis of BLin-UCB

Step 3: argue that we make progress whenever π_{f_t} is not good…

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

$$\leq \sum_{h=0}^{H-1} \sqrt{4T\varepsilon_{gen}^2 + \lambda B_W^2} \left\| X_h(f_t) \right\|_{\Sigma_{t,h}^{-1}}$$

If $V^\star(s_0) - V^{\pi_{f_t}}(s_0) \geq \epsilon$,

Then, we know that $\exists h$, such that $\|X_h(f_t)\|_{\Sigma_{t,h}^{-1}} \geq \epsilon / \sqrt{4T\varepsilon_{gen}^2 + \lambda B_W^2}$

Which means that this new vector $X_h(f_t)$ is "different" from previous "data" $X_h(f_0), \ldots, X_h(f_{t-1})$

i.e., we explore a bit in a d dim space…

Analysis of BLin-UCB

Step 3: argue that we make progress whenever π_{f_t} is not good…

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

$$\leq \sum_{h=0}^{H-1} \sqrt{4T\varepsilon_{gen}^2 + \lambda B_W^2} \; \| X_h(f_t) \|_{\Sigma_{t,h}^{-1}}$$

If $V^\star(s_0) - V^{\pi_{f_t}}(s_0) \geq \epsilon$,

Then, we know that $\exists h$, such that $\|X_h(f_t)\|_{\Sigma_{t,h}^{-1}} \geq \epsilon / \sqrt{4T\varepsilon_{gen}^2 + \lambda B_W^2}$

Which means that this new vector $X_h(f_t)$ is ``different'' from previous ``data'' $X_h(f_0), \ldots, X_h(f_{t-1})$

i.e., we explore a bit in a d dim space…

(We will complete the proof in HW)

Summary for today

Summary for today

1. The BLin-UCB algorithm:

Optimism driven; analysis uses the standard linear bandit style analysis

Summary for today

1. The BLin-UCB algorithm:

Optimism driven; analysis uses the standard linear bandit style analysis

2. The BLin-UCB has poly sample complexity wrt B-rank

It means that this algorithm works for tabular MDPs, linear bandits, linear Bellman-completion, LQRs, Linear $Q^\star$ & $V^\star$, Low-rank MDP, latent variable MDPs, reactive POMDPs, etc

Starting from Thursday:

RL & Optimization:

How to do gradient ascent in RL?
Can gradient ascent find global optimality, despite RL usually has non-convex objective functions?