

Generalization in Large scale MDPs

Wen Sun

CS 6789: Foundations of Reinforcement Learning

Recap: Bellman error of Q

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

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

Recap: Bellman error of Q

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

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

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

Recap: Bellman error of the associated V functions

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

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

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 $\mathcal{E}(Q^*; f, h) = 0, \forall f$

Recap: Bellman error of the associated V functions

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

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

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

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

Recap: The Q / V-Bellman rank

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

	g	f			
π_f	$\mathcal{E}_{g,f,h}$	$\mathcal{E}_{f,f,h}$			

Rank of this Matrix is defined as Bellman Rank

Recap: The Q / V-Bellman rank

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

	g	f			
π_f	$\mathcal{E}_{g;f,h}$	$\mathcal{E}_{f;f,h}$			

There are two mappings
 $W_h : \mathcal{F} \mapsto \mathbb{R}^d$, $X_h : \mathcal{F} \mapsto \mathbb{R}^d$
 ($d = Q/V$ Bellman-rank)

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

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^* & V^* (captures the Q^* -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., $\text{poly}(H, \text{b-rank}, \ln(|\mathcal{F}|), 1/\epsilon^2)$

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 \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:

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, 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 \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 $s_h, a_h \sim d_h^{\pi_f}, s_{h+1} \sim P_h(\cdot | s_h, a_h)$

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

For V-Bellman rank case:

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:

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

For V-Bellman rank case:

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 $s_h \sim d_h^{\pi_f}$, $a_h \sim U(\mathcal{A})$, $s_{h+1} \sim P_h(\cdot | s_h, a_h)$

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

The Algorithm:

At iteration t :

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

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

The Algorithm:

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:

The Algorithm:

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

The Algorithm:

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

Intuition behind the algorithm:

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

Intuition behind the algorithm:

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

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

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

Intuition behind the algorithm:

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

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

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

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

Intuition behind the algorithm:

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

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

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

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

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

Intuition behind the algorithm:

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

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

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

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

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

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

Intuition behind the algorithm:

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

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

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

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

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

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

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 \mathcal{F} :

Analysis of BLin-UCB

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 $1 - \delta$,

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

Analysis of BLin-UCB

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:

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

Analysis of BLin-UCB

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

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

(# of trajectories used: mHT)

Analysis of BLin-UCB

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

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

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

Analysis of BLin-UCB

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

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

Analysis of BLin-UCB

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

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

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)

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

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 $i < t$:

Analysis of BLin-UCB

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

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

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 $i < t$:

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

Analysis of BLin-UCB

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

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

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 $i < t$:

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

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

Analysis of BLin-UCB

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

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

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 $i < t$:

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

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

Analysis of BLin-UCB

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

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

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 $i < t$:

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

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

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

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:

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

Lemma:

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

Analysis of BLin-UCB

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

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

Lemma:

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

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

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

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

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

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

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

$$h = 2, \dots$$

Analysis of BLin-UCB

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

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

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:

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

Analysis of BLin-UCB

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

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

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

Analysis of BLin-UCB

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

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

Analysis of BLin-UCB

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

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

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

Analysis of BLin-UCB

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

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

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

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

Analysis of BLin-UCB

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

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

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

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

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

Analysis of BLin-UCB

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

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

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

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

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

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

Analysis of BLin-UCB

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

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

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

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

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

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

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

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

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

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

Analysis of BLin-UCB

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

$$\begin{aligned} V^\star(s_0) - V^{\pi_{f_t}}(s_0) &\leq \sum_{h=0}^{H-1} \left\| W_h(f_t) \right\|_{\Sigma_{t,h}} \left\| X_h(f_t) \right\|_{\Sigma_{t,h}^{-1}} \\ &\leq \sum_{h=0}^{H-1} \sqrt{4T\varepsilon_{gen}^2 + \lambda B_W^2} \left\| X_h(f_t) \right\|_{\Sigma_{t,h}^{-1}} \end{aligned}$$

Analysis of BLin-UCB

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

$$\begin{aligned} V^\star(s_0) - V^{\pi_{f_t}}(s_0) &\leq \sum_{h=0}^{H-1} \left\| W_h(f_t) \right\|_{\Sigma_{t,h}} \left\| X_h(f_t) \right\|_{\Sigma_{t,h}^{-1}} \\ &\leq \sum_{h=0}^{H-1} \sqrt{4T\epsilon_{gen}^2 + \lambda B_W^2} \left\| X_h(f_t) \right\|_{\Sigma_{t,h}^{-1}} \end{aligned}$$

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...

$$\begin{aligned} V^\star(s_0) - V^{\pi_{f_t}}(s_0) &\leq \sum_{h=0}^{H-1} \left\| W_h(f_t) \right\|_{\Sigma_{t,h}} \left\| X_h(f_t) \right\|_{\Sigma_{t,h}^{-1}} \\ &\leq \sum_{h=0}^{H-1} \sqrt{4T\epsilon_{gen}^2 + \lambda B_W^2} \left\| X_h(f_t) \right\|_{\Sigma_{t,h}^{-1}} \end{aligned}$$

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

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

Analysis of BLin-UCB

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

$$\begin{aligned} V^\star(s_0) - V^{\pi_{f_t}}(s_0) &\leq \sum_{h=0}^{H-1} \left\| W_h(f_t) \right\|_{\Sigma_{t,h}} \left\| X_h(f_t) \right\|_{\Sigma_{t,h}^{-1}} \\ &\leq \sum_{h=0}^{H-1} \sqrt{4T\epsilon_{gen}^2 + \lambda B_W^2} \left\| X_h(f_t) \right\|_{\Sigma_{t,h}^{-1}} \end{aligned}$$

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

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

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

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...

$$\begin{aligned} V^\star(s_0) - V^{\pi_{f_t}}(s_0) &\leq \sum_{h=0}^{H-1} \left\| W_h(f_t) \right\|_{\Sigma_{t,h}} \left\| X_h(f_t) \right\|_{\Sigma_{t,h}^{-1}} \\ &\leq \sum_{h=0}^{H-1} \sqrt{4T\epsilon_{gen}^2 + \lambda B_W^2} \left\| X_h(f_t) \right\|_{\Sigma_{t,h}^{-1}} \end{aligned}$$

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

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

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

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^* & V^* , 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?