

Reinforcement Learning II: Case of Large State Space

Huaning Liu

¹Department of Statistics
University of Illinois at Urbana-Champaign

RL { Intro/Tabular (MDP)
General

Table of Contents

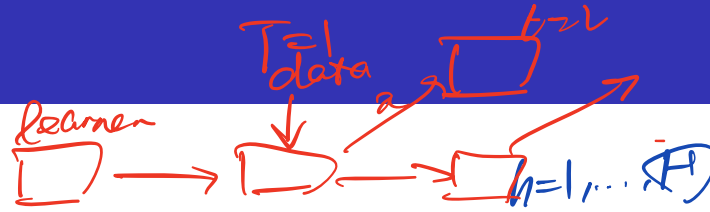
- 1 Discussion on Realizability Assumption
- 2 Linear Function Approximation
- 3 Bellman Rank

Table of Contents

- 1 Discussion on Realizability Assumption
- 2 Linear Function Approximation
- 3 Bellman Rank

Refresher

Recall in Chapter 5 (Intro RL):



- We consider the MDP setting $M = \left\{ \mathcal{S}, \mathcal{A}, \left\{ P_h^M \right\}_{h=1}^H, \left\{ R_h^M \right\}_{h=1}^H, d_1 \right\}$

- $P_h^M : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ is transition probability at h
- $R_h^M : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathbb{R})$ is reward distribution at h
- $d_1 \in \Delta(\mathcal{S}_1)$ is initial state distribution
- \mathcal{S} (resp. \mathcal{A}) is state space (resp. action space)

- The online regime interacts with unknown MDP M^* for T episodes, for each $t \in [T]$, under selected policy $\pi^t \in \Pi_{RNS}$, we observe the trajectory (data) $\tau^t = \{(s_1^t, a_1^t, r_1^t), \dots, (s_H^t, a_H^t, r_H^t)\}$

- Target: minimize the total regret

$$\sum_{t=1}^T \mathbb{E}_{\pi^t \sim p_t} \left[f^{M^*}(\pi^{M^*}) - f^{M^*}(\pi^t) \right], \text{ where } f^M(\pi) := \mathbb{E}^{M, \pi} \left[\sum_{h=1}^H r_h \right]$$

- Regret bound under Tabular MDP derived depends on $|\mathcal{S}|$ and $|\mathcal{A}|$

Refresher II

[1] Chapter 6 introduced a general decision making framework, here we focus back on MDP only; trivially assume $\sum_{h=1}^H r_h \in [0, 1]$,

$$\Pi = \Pi_{RNS} \equiv \{\pi : \pi = (\pi_1, \dots, \pi_H), \pi_h : \mathcal{S} \rightarrow \Delta(\mathcal{A})\}$$

Also recall state value function and state-action value function

$$V_h^{M,\pi}(s) = \mathbb{E}^{M,\pi} \left[\sum_{h'=h}^H r_{h'} \mid s_h = s \right]$$
$$Q_h^{M,\pi}(s, a) = \mathbb{E}^{M,\pi} \left[\sum_{h'=h+1}^H r_{h'} \mid s_h = s, a_h = a \right]$$

Question: When the MDP is no longer tabular (i.e. large $|\mathcal{S}|$), can we still derive an effective regret bound that doesn't depend on this?

Realizability as assumption

$M^* \in \mathcal{M}$

One can consider several types of realizability assumptions in RL

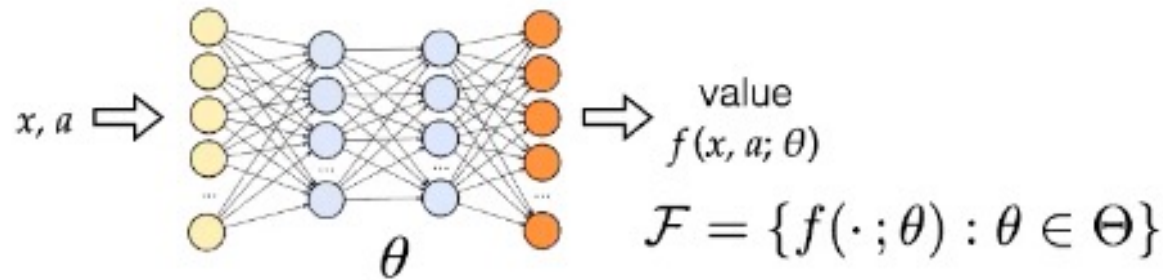
- 1 **Model realizability**: Have access to a model class \mathcal{M} of MDPs that contains the true MDP M^* .
- 2 **Value function realizability**: Have access to a class \mathcal{Q} of state-action value functions (Q functions) that contains the optimal function $Q^{M^*,*}$ for the underlying MDP.
- 3 **Policy realizability**: Have access to a class Π of policies that contains the optimal policy π_{M^*} .

Remark: $1 \Rightarrow 2 \Rightarrow 3$

- Ideally when 1 (resp. 2, 3) holds, we'll be able to bound the regret with $|\mathcal{M}|$ (resp. $|\mathcal{Q}|, |\Pi|$). $\log |\mathcal{M}|$
- For example(2), given \mathcal{Q} such that $Q^* \in \mathcal{Q}$, w.p $1 - \delta$, find policy π such that $\mathbb{E}[\text{Reg}] \leq \epsilon$ using $\text{poly}(|\mathcal{A}|, H, \log |\mathcal{Q}|, 1/\epsilon, 1/\delta)$ episodes.

$\Downarrow \mathbb{E}[\text{Reg}] \leq \text{poly}(|\mathcal{A}|, \dots)$

Realizability as assumption II



Claim: With this single assumption, this is not achievable.

Proposition (1)

(Krishnamurthy et al.): For any $S \in \mathbb{N}$ and $H \in \mathbb{N}$, there exists a class of horizon- H MDPs \mathcal{M} with $|\mathcal{S}| = S$, $|\mathcal{A}| = 2$, and $\log |\mathcal{M}| = \log(S)$, yet any algorithm must have

$$\mathbb{E}[\text{Reg}] \gtrsim \sqrt{\min\{S, 2^H\} \cdot T}$$

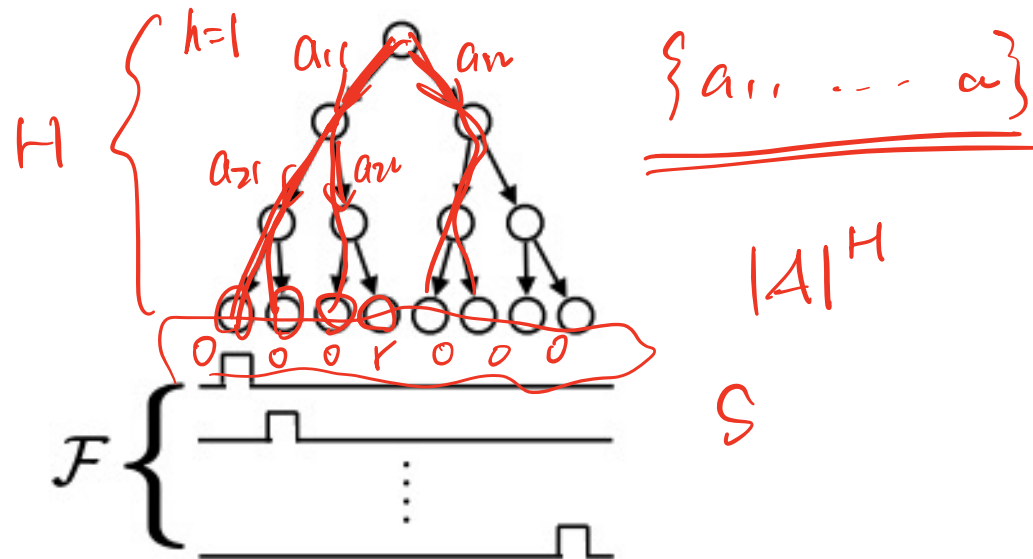
Handwritten notes: "unbounded" and "exp H" with arrows pointing to the 2^H term in the equation.

This implied neither of the three assumption suffices, since the setting considers model realizability already.

Realizability as assumption III

(Quick Proof Sketch)

- [2] Idea: Since $|\mathcal{S}|$ unbounded, the construction use a depth- H complete tree to emulate Multi-armed bandit with $|\mathcal{A}|^H$ arms
- Recall that sample complexity lower bound for MAB is $\frac{\# \text{ arms}}{\varepsilon^2}$ 2^H
- Thus, without any restriction (function approximation), the exploration algorithms have exponential sample complexity - it suffices to show function approximation does not help.

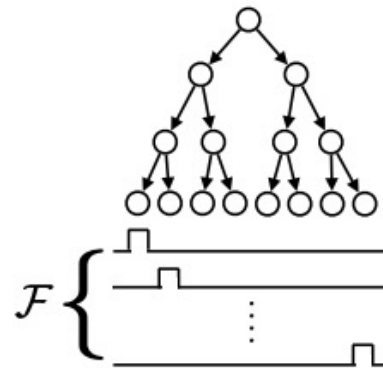


Realizability as assumption IV

(Quick Proof Sketch cont'd)

- By definition, \mathcal{Q} is a collection of Q^* from all MDPs in family (note any state as leaf node could be chosen to be rewarding)
- Smallest possible $|\mathcal{Q}|$ that let \mathcal{Q} realizable will be $\log |\mathcal{Q}| = H \log |\mathcal{A}|$.
- Restricting \mathcal{Q} doesn't really help!

$|\mathcal{A}|^H$



Thus, we would like to explore more possible assumptions in addition to realizability, which could possibly enables 1. extrapolation across state space, 2. determine some effective states that algo mainly learn from

Table of Contents

1 Discussion on Realizability Assumption

2 Linear Function Approximation

3 Bellman Rank

Linear- Q^* model

One intuitive assumption is the linearity of the underlying Q -function wrt true model M

$$Q_h^{M,*}(s, a) = \langle \phi(s, a), \theta_h^M \rangle, \quad \forall h \in [H]$$

- $\phi(s, a) \in \mathbb{B}_2^d(1)$ is a known feature (analogue to structured bandit ch)
- $\theta_h^M \in \mathbb{B}_2^d(1)$ is unknown parameter

Or equivalently assume $Q^{M,*} \in \mathcal{Q}$, where

$$\mathcal{Q} = \left\{ Q_h(s, a) = \langle \phi(s, a), \theta_h \rangle \mid \theta_h \in \mathbb{B}_2^d(1) \forall h \right\}$$

Unfortunately, the regret is still lower bounded by exponential term.

Proposition (2)

(Weisz et al.) For any $d \in \mathbb{N}$ and $H \in \mathbb{N}$ sufficiently large, any algorithm for the Linear- Q^* model must have

$$\mathbb{E}[\text{Reg}] \gtrsim \min \left\{ 2^{\Omega(d)}, 2^{\Omega(H)} \right\}$$

Low-Rank MDP

Another proper but bit stronger assumption is the assertion of **linearity behind transition probabilities** $\phi(s, a) \in \mathbb{R}^d$

$$P_h^M(s' | s, a) = \langle \phi(s, a), \mu_h^M(s') \rangle, \quad \text{and} \quad \mathbb{E}[r_h | s, a] = \langle \phi(s, a), w_h^M \rangle$$

- $\phi(s, a) \in B_2^d(1)$ known feature map, $\mu_h^M(s') \in \mathbb{R}^d$ unknown feature
- $w_h^M \in B_2^d(\sqrt{d})$ unknown parameter

For simplicity further assume $\left\| \sum_{s' \in \mathcal{S}} \mu_h^M(s') \right\| \leq \sqrt{d}$ and stepwise & cumulative reward in $[0, 1]$.

Remark: The transition matrix has rank **at most d** , regardless of $|\mathcal{A}|$ and $|\mathcal{S}|$. That's namely why the structure **low-rank MDP**.

Remark: It generalize tabular MDP, where $\phi(s, a) = \mathbf{e}_{s,a}$ and $(\mu_h(s'))_{s,a} = P_h^M(s' | s, a)$.

Property of Low-Rank MDP

Lemma (1)

(Linearity of Bellman backup) For any low-rank MDP $M \in \mathcal{M}$ and any $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ and any $h \in [H]$, the Bellman operator is linear in ϕ :

$$[\mathcal{T}_h^M Q](s, a) = \langle \phi(s, a), \theta_Q^M \rangle \quad (1)$$

for some $\theta_Q^M \in \mathbb{R}^d$. In particular, this implies that for any policy $\pi = (\pi_1, \dots, \pi_H)$, functions $Q_h^{M, \pi}$ are linear in ϕ for every h . Finally, for $Q : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$, it holds that $\|\theta_Q^M\| \leq 2\sqrt{d}$

Importance: 1. Connect linear structure w/ optimality 2. Implies low-rank being stronger than linear- Q^*

Proof...

Proof (Lemma 1)

First observe

$$\begin{aligned}
 [T_h^M Q](s, a) &= \underbrace{\langle \phi(s, a), w_h^M \rangle}_{\mathbb{E}[r]} + \underbrace{\sum_{s'} P_h^M(s' | s, a) \max_{a'} Q(s', a')}_{+ \max Q} \\
 &= \langle \phi(s, a), w_h^M \rangle + \sum_{s'} \langle \phi(s, a), \mu_h^M(s') \rangle \max_{a'} Q(s', a') \\
 &= \left\langle \phi(s, a), w_h^M + \sum_{s'} \mu_h^M(s') \max_{a'} Q(s', a') \right\rangle.
 \end{aligned}$$

tabular MDP
 $\hat{P}(s'|s, a)$
 $s' \in S$

$(s, a) \mapsto \left(\begin{array}{c} \\ \end{array} \right)$
 $(s, a) \mapsto$

Also

$$\|\theta_Q^M\| \leq \|w_h^M\| + \left\| \sum_{s'} \mu_h^M(s') Q(s') \right\| \leq 2\sqrt{d}$$

LSVI-UCB

- Idea: Construct optimism that $\bar{Q}_h^t(s, a) \geq Q_h^{M, \star}(s, a)$ for all s, a, h .
- Challenge: Unlike UCB-VI, empirically estimate trans prob no longer make sense given the large $|\mathcal{S}|$ (nor feasible regarding unknown μ^M).
- Trick: Take adv. of Lemma 1, use Least Square to regress $Y = r_h + \max_a \bar{Q}_{h+1}^t(s_{h+1}, a)$ onto $X = \phi(s_h, a_h)$.

LSVI-UCB

Input: $R, \rho > 0$

for $t = 1, \dots, T$ do

Let $\bar{Q}_{H+1}^t \equiv 0$.

for $h = H, \dots, 1$ do

Compute least-squares estimator

$$\hat{\theta}_h^t = \arg \min_{\theta \in \mathcal{B}_2^d(\rho)} \sum_{i < t} \left(\langle \phi(s_h^i, a_h^i), \theta \rangle - r_h^i - \max_a \bar{Q}_{h+1}^t(s_{h+1}^i, a) \right)^2,$$

and let $\hat{Q}_h^t(s, a) := \langle \phi(s, a), \hat{\theta}_h^t \rangle$.

Define

$$\Sigma_h^t = \sum_{i < t} \phi(s_h^i, a_h^i) \phi(s_h^i, a_h^i)^\top + I.$$

Compute bonus:

$$b_{h,\delta}^t(s, a) = \sqrt{R} \|\phi(s, a)\|_{(\Sigma_h^t)^{-1}}.$$

Compute optimistic value function:

$$\bar{Q}_h^t(s, a) = \left\{ \hat{Q}_h^t(s, a) + b_{h,\delta}^t(s, a) \right\} \wedge 1.$$

Set $\bar{V}_h^t(s) = \max_{a \in \mathcal{A}} \bar{Q}_h^t(s, a)$ and $\hat{\pi}_h^t(s) = \arg \max_{a \in \mathcal{A}} \bar{Q}_h^t(s, a)$.

Collect trajectory $(s_1^t, a_1^t, r_1^t), \dots, (s_H^t, a_H^t, r_H^t)$ according to $\hat{\pi}^t$.

\uparrow
 ρ

tabular:

$$\sqrt{\frac{\dots}{u(s, a)}} \approx \frac{e_{s, a} \cdot e_{s, a}^\top}{\dots}$$

LSVI-UCB II

Remark: here \bar{Q}_h^t belongs to class $\mathcal{Q} :=$

$$\left\{ (s, a) \mapsto \left\{ \langle \theta, \phi(s, a) \rangle + \sqrt{R} \|\phi(s, a)\|_{(\Sigma)^{-1}} \right\} \wedge 1 : \|\theta\| \leq 2\sqrt{d}, \sigma_{\min}(\Sigma) \geq 1 \right\}$$

Proposition (3)

(Regret bound) If any $\delta > 0$, if we set $R = c \cdot d^2 \log(HT/\delta)$ for a sufficiently large numerical constant c and $\rho = 2\sqrt{d}$, LSVI-UCB has that with probability at least $1 - \delta$,

LSI

$$\text{Reg} \lesssim H \sqrt{d^3 \cdot T \log(HT/\delta)}$$

Proof idea

- 1 Regression: closure between $\left[T_h^M \bar{Q}_{h+1}^t \right] (s, a)$ and $\hat{Q}_h^t(s, a)$.
- 2 Ensure optimism (and feasibility of Q)
- 3 Analysis of regret

Lemma (2)

Let \mathcal{G} be an abstract set with $|\mathcal{G}| < \infty$. Let $x_1, \dots, x_T \in \mathcal{X}$ be fixed, and for each $g \in \mathcal{G}$, let $y_1(g), \dots, y_T(g) \in \mathbb{R}$ be 1-subGaussian outcomes satisfying $\mathbb{E}[y_i(g) \mid x_i] = f_g(x_i)$ for $f_g \in \mathcal{F} \subseteq \{f : \mathcal{X} \rightarrow \mathbb{R}\}$.^a In addition, assume that $y_1(g), \dots, y_T(g)$ are conditionally independent given x_1, \dots, x_T . For any latent $g \in \mathcal{G}$, define the least-squares solution

$$\hat{f}_g \in \arg \min_{f \in \mathcal{F}} \sum_{i=1}^T (y_i(g) - f(x_i))^2.$$

With probability at least $1 - \delta$, simultaneously for all $g \in \mathcal{G}$,

$$\sum_{i=1}^T (\hat{f}_g(x_i) - f_g(x_i))^2 \lesssim \log(|\mathcal{F}| |\mathcal{G}| / \delta).$$

Lemma (3) ~~✗~~

With probability at least $1 - \delta$, we have that for all t and h ,

$$\sum_{i < t} \left(\hat{Q}_h^t(s_h^i, a_h^i) - \left[\mathcal{T}_h^M \bar{Q}_{h+1}^t \right] (s_h^i, a_h^i) \right)^2 \lesssim d^2 \log(HT/\delta)$$

Proof (Lemma 2)

- Denote empirical norm $\|f\|_T^2 = \frac{1}{T} \sum_{i=1}^T f(x_i)^2$
- Fix $g \in \mathcal{G}$, first observe $\|Y_g - \hat{f}_g\|_T^2 \leq \|Y_g - f_g\|_T^2$ by optimality of \hat{f}_g .
- This implies $\|\hat{f}_g - f_g\|_T^2 \leq 2 \langle Y_g - f_g, \hat{f}_g - f_g \rangle_T$, it then follows

$$\|\hat{f}_g - f_g\|_T \leq 2 \max_{f \in \mathcal{F}} \left\langle Y_g - f_g, \frac{f - f_g}{\|f - f_g\|_T} \right\rangle_T$$
- Note $Y_g - f_g$ being centered 1-subGaussian, and right term is vector with Euclidean length \sqrt{T} (why?)
- Thus algebraic conclusion regarding sub-Gaussian vector tells us

$$RHS \leq C \sqrt{\log(|\mathcal{F}|/\delta)/T}$$

Take union bound across \mathcal{G} yields the result.

Proof (Lemma 3)

Analogously use lemma 2: fix h and t , consider data $(s_h^i, a_h^i, s_{h+1}^i, r_h^i)$

- Q matches \mathcal{G}
- $r_h^i + \max_a Q(s_{h+1}^i, a)$ matches $y_i(Q)$
- $\phi(s_h^i, a_h^i)$ matches x^i
- Again apply Lemma 1, we check

$$\begin{aligned}\mathbb{E}[y_i(Q) | x_i] &= \mathbb{E}^M \left[r_h^i + \max_a Q(s_{h+1}^i, a) \mid s_h^i, a_h^i \right] \\ &= [\mathcal{T}_h^M Q](s_h^i, a_h^i) = \langle \phi(s_h^i, a_h^i), \theta_Q^M \rangle \equiv f_g(x_i)\end{aligned}$$

- Thus the regression will be well-specified as long as we choose

$$\log \frac{|\mathcal{Q}| |\mathcal{F}| \dots}{2}$$

$$\mathcal{F} = \{ \phi(s, a) \mapsto \langle \phi(s, a), \theta \rangle : \|\theta\| \leq 2\sqrt{d} \}$$

- A proper choice of scale ϵ achieves ϵ -discretized classes towards the covering of \mathcal{Q} and \mathcal{F} , whose order can be shown $\tilde{O}(d)$ and $\tilde{O}(d^2)$ resp. Union bound over T and H yields the result.

Optimism

Generalize the result from data $\{(s_h^i, a_h^i)\}_{i < t}$ to arbitrary (s, a) pair.

Lemma (4)

Whenever the event in Lemma 3 occurs, we have that for all (s, a, h) and $t \in [T]$,

$$\left| \widehat{Q}_h^t(s, a) - [T_h^M \bar{Q}_{h+1}^t](s, a) \right| \lesssim \sqrt{d^2 \log(HT/\delta) \cdot \|\phi(s, a)\| (\Sigma_h^t)^{-1}} = b_{h,\delta}^t(s, a). \quad (2)$$

and

$$\bar{Q}_h^t(s, a) \geq Q_h^{M,*}(s, a). \quad (3)$$

Proof (Lemma 4)

Proof for (2), (using Lemma 1) use the linearity of Bellman backup.
Algebraically

$$\begin{aligned}
 \left| \widehat{Q}_h^t(s, a) - \left[\mathcal{T}_h^M \bar{Q}_{h+1}^t \right] (s, a) \right| &= \left| \langle \phi(s, a), \widehat{\theta}_h^t - \theta_h^t \rangle \right| \\
 \underbrace{\left| \langle \phi(s, a), \widehat{\theta}_h^t \rangle \right|}_{\langle \phi(s, a), \widehat{\theta}_h^t \rangle} - \underbrace{\left| \langle \phi(s, a), \theta_h^t \rangle \right|}_{\langle \phi(s, a), \theta_h^t \rangle} &\rightarrow \left| \langle (\Sigma_h^t)^{-1/2} \phi(s, a), (\Sigma_h^t)^{1/2} (\widehat{\theta}_h^t - \theta_h^t) \rangle \right| \\
 &\leq \|\phi(s, a)\|_{(\Sigma_h^t)^{-1}} \cdot \left\| \widehat{\theta}_h^t - \theta_h^t \right\|_{\Sigma_h^t} \quad \leftarrow |AB| \leq \|A\| \cdot \|B\|
 \end{aligned}$$

Decompose the second term² gives

$$\begin{aligned}
 \left\| \widehat{\theta}_h^t - \theta_h^t \right\|_{\Sigma_h^t}^2 &= \left(\widehat{\theta}_h^t - \theta_h^t \right)^T \left(\sum_{i < t} \phi(s_i^i, a_i^i) \phi(s_i^i, a_i^i)^T + \mathcal{I} \right) \left(\widehat{\theta}_h^t - \theta_h^t \right) \\
 &= \sum_{i < t} \left(\widehat{Q}_h^t(s_i^i, a_i^i) - \left[\mathcal{T}_h^M \bar{Q}_{h+1}^t \right] (s_i^i, a_i^i) \right)^2 + \left\| \widehat{\theta}_h^t - \theta_h^t \right\|_{\Sigma_h^t}^2 \\
 &\lesssim d^2 \log(HT/\delta)
 \end{aligned}$$

Proof (Lemma 4, cont.)

Proof for (3), use result from (2) inductively

- Base case: $\bar{V}_{H+1}^t = V_{H+1}^{M,*} \equiv 0$
- Inductively: Suppose $\bar{V}_{h+1}^t \geq V_{h+1}^{M,*}$, by monotonicity of Bellman Operator $\mathcal{T}_h^M \bar{V}_{h+1}^t \geq \mathcal{T}_h^M V_{h+1}^{M,*} = Q_h^{M,*}$. Thus

$$\begin{aligned}\hat{Q}_h^t &= \hat{Q}_h^t - \mathcal{T}_h^M \bar{V}_{h+1} + \mathcal{T}_h^M \bar{V}_{h+1} \\ &\geq \hat{Q}_h^t - \mathcal{T}_h^M \bar{V}_{h+1} + Q_h^{M,*} \\ &\geq -b_{h,\delta}^t + Q_h^{M,*}\end{aligned}$$

This directly implies $\hat{Q}_h^t \geq Q_h^{M,*}$

Proof (Finishing)

Denote the true model M , then at time t

$$\begin{aligned} f^M(\pi_M) - f^M(\hat{\pi}^t) &\leq \mathbb{E}_{s_1 \sim d_1} [\bar{V}_1^t(s_1)] - f^M(\hat{\pi}^t) \quad (\text{Optimistic}) \\ &= \sum_{h=1}^H \mathbb{E}^{M, \hat{\pi}^t} [\bar{Q}_h^t(s_h, a_h) - [\mathcal{T}_h^M \bar{Q}_{h+1}^t](s_h, a_h)] \\ &\lesssim \sqrt{R} \sum_{h=1}^H \mathbb{E}^{M, \hat{\pi}^t} [\|\phi(s_h, a_h)\| (\Sigma_h^t)^{-1}] \quad \begin{array}{l} \downarrow \text{lemma 4} \\ (\text{Bellman} \\ \text{Decomposition}) \end{array} \end{aligned}$$

The regret follows

$$\mathbf{Reg} \leq \sqrt{R} \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}^{M, \hat{\pi}^t} [\|\phi(s_h, a_h)\| (\Sigma_h^t)^{-1}]$$

Apply Hoeffding (w/ union bound) and elliptic lemma gives the desired result. This completes the proof.

Table of Contents

1 Discussion on Realizability Assumption

2 Linear Function Approximation *(low-rank MDP)*

3 Bellman Rank

Intuition of Bellman Rank

We try to relax the linear structure behind \mathcal{Q} in low-rank MDP.

- **Intuition:** Fix $h \in [H]$, for any function $Q \in \mathcal{Q}$ and $\pi \in \Pi$, decompose the *Bellman residual* wrt Q_h under π to be the linear combination between some embedding of π and Q .

- **Example:** In low-rank MDP, the Bellman residual writes

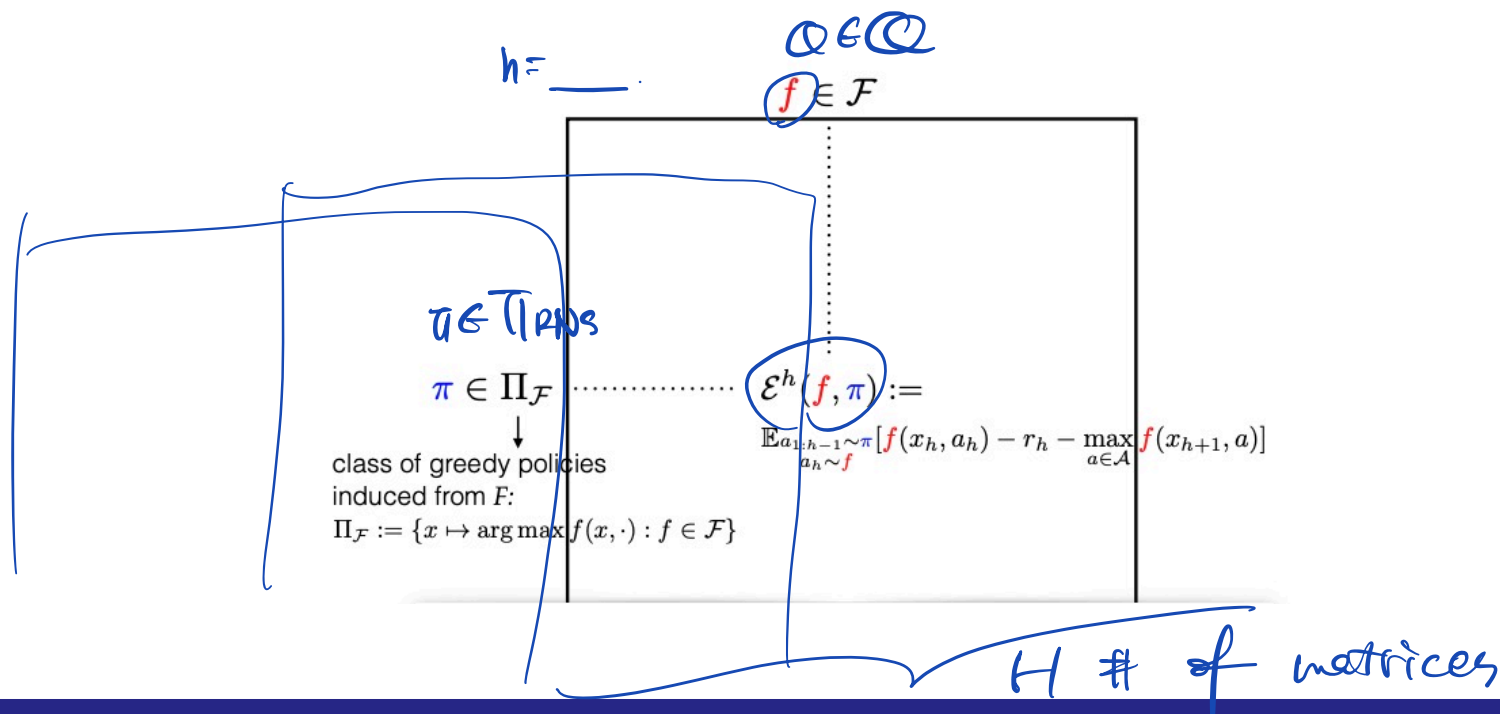
$$\begin{aligned} & \mathbb{E}^{M, \pi} \left[Q_h(s_h, a_h) - r_h - \max_a Q_{h+1}(s_{h+1}, a) \right] \\ &= \left\langle \mathbb{E}^{M, \pi} [\phi(s_h, a_h)], \theta_h^Q - w_h^M - \tilde{\theta}_h^{M, Q} \right\rangle = \left\langle X_h^M(\pi), W_h^M(Q) \right\rangle \end{aligned}$$

- View the Bellman residual as $\Pi \times \mathcal{Q}$ matrix $\mathcal{E}_h(\cdot, \cdot) \in \mathbb{R}^{\Pi \times \mathcal{Q}}$ with

$$\mathcal{E}_h(\pi, Q) := \mathbb{E}^{M, \pi} \left[Q_h(s_h, a_h) - \left(r_h + \max_a Q_{h+1}(s_{h+1}, a) \right) \right]$$

In the example, it's immediate $\text{rank}(\mathcal{E}_h(\cdot, \cdot)) \leq d$.

Bellman Rank



Definition

For an MDP M with value function class \mathcal{Q} and policy class Π , the Bellman rank is defined as

$$d_B(M) = \max_{h \in [H]} \text{rank} \left(\{ \mathcal{E}_h(\pi, Q) \}_{\pi \in \Pi, Q \in \mathcal{Q}} \right)$$

Remark: Regime of low Bellman rank implicitly control the distribution shifts in MDP M .

(7.3.1)

Examples of Bellman Rank I

(7.3.3)

There're several examples of models with low Bellman Rank.

Tabular MDP: M with $|\mathcal{S}| \leq S$ and $|\mathcal{A}| \leq A$

$$\begin{aligned} \mathcal{E}_h(\pi, Q) &= \mathbb{E}^{M, \pi} \left[Q_h(s_h, a_h) - \left(r_h + \max_a Q_{h+1}(s_{h+1}, a) \right) \right] \\ &= \sum_{s, a} d_h^{M, \pi}(s, a) \mathbb{E}^M \left[Q_h(s, a) - \left(r_h + \max_{a'} Q_{h+1}(s_{h+1}, a') \right) \mid s_h = s, a_h = a \right] \end{aligned}$$

\mathbb{R}^{SA}
 \downarrow
 \downarrow
 $\langle X, W \rangle$

Take $X_h^M(\pi) \in \mathbb{R}^{SA}$ as the first term and $W_h^M(Q) \in \mathbb{R}^{SA}$ as the second yields $\mathcal{E}_h(\pi, Q) = \langle X_h^M(\pi), W_h^M(Q) \rangle$. It follows

$$d_B(M) \leq SA$$

Examples of Bellman Rank II

Low Occupancy Complexity: For MDP M , there exists a feature map $\phi^M(s, a) \in \mathbb{R}^d$ such that for all π , there exists $\theta_h^{M, \pi} \in \mathbb{R}^d$ such that

$$d_h^{M, \pi}(s, a) = \langle \phi^M(s, a), \theta_h^{M, \pi} \rangle, \text{ both terms unknown}$$

Use similar trick we have

$$\begin{aligned} \mathcal{E}_h(\pi, Q) &= \mathbb{E}^{M, \pi} \left[Q_h(s_h, a_h) - \left(r_h + \max_a Q_{h+1}(s_{h+1}, a) \right) \right] \\ &= \sum_{s, a} d_h^{M, \pi}(s, a) \mathbb{E}^M \left[Q_h(s, a) - \left(r_h + \max_{a'} Q_{h+1}(s_{h+1}, a') \right) \mid s_h = s, a_h = a \right] \\ &= \sum_{s, a} \langle \phi^M(s, a), \theta_h^{M, \pi} \rangle \mathbb{E}^M \left[Q_h(s, a) - \left(r_h + \max_{a'} Q_{h+1}(s_{h+1}, a') \right) \mid s_h = s, a_h = a \right] \\ &= \left\langle \theta_h^{M, \pi}, \sum_{s, a} \phi^M(s, a) \mathbb{E}^M \left[Q_h(s, a) - \left(r_h + \max_{a'} Q_{h+1}(s_{h+1}, a') \right) \mid s_h = s, a_h = a \right] \right\rangle. \end{aligned}$$

Then clearly $d_B(M) \leq d$. Comment: This model generalize low-rank MDP and tabular MDP, for allowing non-linear function approximation as long as definition above holds.

Examples of Bellman Rank III

Linear Quadratic Regulator (LQR): $\mathcal{S} = \mathcal{A} = \mathbb{R}^d$

The dynamics are assumed $s_{h+1} = A^M s_h + B^M a_h + \zeta_h$ where $\zeta_h \sim \mathcal{N}(0, I)$, and $s_1 \sim \mathcal{N}(0, I)$, and $r_h = -s_h^\top Q^M s_h - a_h^\top R^M a_h$ for some (usually known) matrices $Q^M, R^M \succeq 0$. cost

A classic result gives linear optimal control and quadratic value function

$$\pi_{M,h}(s) = K_h^M s \text{ and } Q^{M,*}(s, a) = (s, a)^\top P_h^M(s, a)$$

It can then be shown $d_B(M) \leq d^2 + 1$. (idea: match π with the quadratic structure of Q towards Bellman residual)

BiLinUCB Algorithm (Intro)

- Setting: MDP with low Bellman rank + realizability $Q^{M^*,*} \in \mathcal{Q}$
- Try derive **PAC** ("Probably Approximately Correct") guarantee (in replacement of regret), measured by $f^{M^*}(\pi_{M^*}) - f^{M^*}(\hat{\pi})$ i.e. concern only the *final performance*
- Learning regime: Want to ensure $f^{M^*}(\pi_{M^*}) - f^{M^*}(\hat{\pi}) \leq \epsilon$ for some $\epsilon \ll 1$ using $\text{poly}(\frac{1}{\epsilon})$ # episodes
- ✓ Remark: This is **easier**: for some algo with $\mathbb{E}[\text{Reg}] \lesssim \sqrt{CT} \implies$ PAC with $O(\frac{C}{\epsilon^2})$ episodes; the other way gives $\mathbb{E}[\text{Reg}] \lesssim C^{1/3} T^{2/3}$.
 (PAC with $O(\frac{C}{\epsilon^2})$ episodes)



"Online - to - Batch" conversation: $\mathbb{E}[\epsilon c f^{\uparrow}] \leq \frac{1}{T} \mathbb{E}[\text{Reg}] \leq \sqrt{\frac{C}{T}} \sim \epsilon$

$\Leftrightarrow T \sim O(\frac{C}{\epsilon^2})$

BiLinUCB Algorithm

Algorithm intuition

- Takes K iterations, each consists n episodes *(batch)*
- Maintains a confidence set $Q^k \subseteq Q$ where w.h.p $Q^{M^*,*} \in Q^k$
- For each iteration:
 - 1 Compute optimistic-on-average value function

(b_n) $\bar{Q}_h \approx Q^*$ $\forall h$ $Q^k = \arg \max_{Q \in Q^k} \mathbb{E}_{s_1 \sim d_1} [Q_1(s_1, \pi_Q(s_1))]$

and the corresponding policy $\pi^k := \pi_{Q^k}$ (That's why here we "search for optimism" instead of "construct optimism")

- 2 Use *(current batch)* n episodes to estimate Bellman residual wrt π^k for all $Q \in Q$, namely $\{\hat{\mathcal{E}}_h^k(Q)\}_{h \in [H]}$
- 3 Then with some threshold *fixed*, choose those $Q' \in Q$ with low estimated Bellman residual, since optimally $Q^{M^*,*}$ has BR of 0.

BiLinUCB Algorithm

Statement (pseudo)

BiLinUCB

Input: $\beta > 0$, iteration count $K \in \mathbb{N}$, batch size $n \in \mathbb{N}$.

$\mathcal{Q}^1 \leftarrow \mathcal{Q}$.

for iteration $k = 1, \dots, K$ do

 Compute optimistic value function:

$$Q^k = \arg \max_{Q \in \mathcal{Q}^k} \mathbb{E}_{s_1 \sim d_1} [Q_1(s_1, \pi_Q(s_1))].$$

 and let $\pi^k := \pi_{Q^k}$.

 for $l = 1, \dots, n$ do

 Execute π^k for an episode and observe trajectory $(s_1^{k,l}, a_1^{k,l}, r_1^{k,l}), \dots, (s_H^{k,l}, a_H^{k,l}, r_H^{k,l})$.

 Compute confidence set

$$\mathcal{Q}^{k+1} = \left\{ Q \in \mathcal{Q} \mid \sum_{i \leq k} (\hat{\mathcal{E}}_h^i(Q))^2 \leq \beta \quad \forall h \in [H] \right\}, \quad (7.26)$$

 where

$$\hat{\mathcal{E}}_h^i(Q) := \frac{1}{n} \sum_{l=1}^n \left(Q_h(s_h^{i,l}, a_h^{i,l}) - r_h^{i,l} - \max_{a \in \mathcal{A}} Q_{h+1}(s_{h+1}^{i,l}, a) \right).$$

Let $\hat{k} = \arg \max_{k \in [K]} \hat{V}^k$, where $\hat{V}^k := \frac{1}{n} \sum_{l=1}^n \sum_{h=1}^H r_h^{k,l}$.

Return $\hat{\pi} = \pi^{\hat{k}}$.

Main Result from BiLinUCB

Proposition (4)

Suppose that M^* has Bellman rank d and $Q^{M^*,*} \in \mathcal{Q}$. For any $\varepsilon > 0$ and $\delta > 0$, if we set $n \gtrsim \frac{H^3 d \log(|\mathcal{Q}|/\delta)}{\varepsilon^2}$, $K \gtrsim Hd \log(1 + n/d)$, and $\beta \propto c \cdot K \frac{\log |\mathcal{Q}| + \log(HK/\delta)}{n}$, then BiLinUCB learns a policy $\hat{\pi}$ such

$$f^{M^*}(\pi_{M^*}) - f^{M^*}(\hat{\pi}) \leq \varepsilon$$

with probability at least $1 - \delta$, and does so using

$$\tilde{O}\left(\frac{H^4 d^2 \log(|\mathcal{Q}|/\delta)}{\varepsilon^2}\right) \# \text{ episodes.}$$

Equivalently

$$\text{Reg} \leq \tilde{O}\left(\left(H^4 d^2 \log(|\mathcal{Q}|/\delta)\right)^{1/3} \cdot T^{2/3}\right)$$

episodes.

Q: What's the new trick here? How Bellman rank comes into play?

Proof I

For simplicity assume $\|X_h^{M^*}(\pi)\|, \|W_h^{M^*}(Q)\|_2 \leq 1$. The general pipeline is

- 1 Justify the validity of Q^k each time of construction
- 2 Verify the optimism of constructed value function
- 3 Back to the main result

Lemma (5)

For any $\delta > 0$, if we set $\beta = c \cdot K \frac{\log |Q| + \log(HK/\delta)}{n}$, where $c > 0$ is sufficiently large absolute constant, then with probability at least $1 - \delta$, for all $k \in [K]$: 1. All $Q \in Q^k$ have

$$\sum_{i < k} \left(\mathcal{E}_h(\pi^i, Q) \right)^2 \lesssim \beta \quad \forall h \in [H].$$

2. $Q^{M^*,*} \in Q^k$. ✓

Proof (Lemma 5)

- First by Hoeffding and Union bound, it's direct w.p. $1 - \delta$, for all $k \in [K]$, $h \in [H]$, and $Q \in \mathcal{Q}$,

$$\left| \widehat{\mathcal{E}}_h^k(Q) - \mathcal{E}_h(\pi^k, Q) \right| \leq C \cdot \sqrt{\frac{\log(|\mathcal{Q}|HK/\delta)}{n}}$$

- Then by AM-GM inequality for all $k \in [K]$

$$\sum_{i < k} (\mathcal{E}_h(\pi^i, Q))^2 \leq 2 \sum_{i < k} (\widehat{\mathcal{E}}_h^i(Q))^2 + 2 \sum_{i < k} (\mathcal{E}_h(\pi^i, Q) - \widehat{\mathcal{E}}_h^i(Q))^2$$

$\leq \beta \sum_{Q \in \mathcal{Q}} \sum_{i < k} (\widehat{\mathcal{E}}_h^i(Q))^2 + \beta \sum_{i < k} (\mathcal{E}_h(\pi^i, Q))^2$

(1) is then immediate by definition of Q^k

- Similarly observe for all $k, h, Q \in \mathcal{Q}$

$$\sum_{i < k} (\widehat{\mathcal{E}}_h^i(Q))^2 \leq 2 \sum_{i < k} (\mathcal{E}_h(\pi^i, Q))^2 + 2 \sum_{i < k} (\mathcal{E}_h(\pi^i, Q) - \widehat{\mathcal{E}}_h^i(Q))^2$$

$= 0 + \beta \sum_{i < k} (\mathcal{E}_h(\pi^i, Q))^2$

(2) is then immediate by noting $\mathcal{E}_h(\pi, Q^{M^*, *}) = 0 \quad \forall \pi$

Justify optimism

Lemma (6)

Whenever the event in Lemma(5) occurs, the following properties hold:

1. Define

$$\Sigma_h^k = \sum_{i < k} X_h^{M^*}(\pi^i) X_h^{M^*}(\pi^i)^\top.$$

For all $k \in [K]$, all $Q \in \mathcal{Q}^k$ satisfy

$$\left\| W_h^{M^*}(Q) \right\|_{\Sigma_h^k}^2 \lesssim \beta. \quad \checkmark$$

2. For all k , Q^k is optimistic in the sense that

$$\mathbb{E}_{s_1 \sim d_1} [Q_1^k(s_1, \pi_{Q^k}(s_1))] \geq \mathbb{E}_{s_1 \sim d_1} [Q_1^{M^*,*}(s_1, \pi_{M^*}(s_1))] = f^{M^*}(\pi_{M^*}) \quad \checkmark$$

Proof (Lemma 6)

- Recall bilinear class property $\exists X_h^{M^*}(\pi), W_h^{M^*}(Q) \in \mathbb{R}^d$ s.t.
 $\mathcal{E}_h(\pi, Q) = \langle X_h^{M^*}(\pi), W_h^{M^*}(Q) \rangle$

- By Lemma(5) and some calculation, we have

$$\|W^{M^*}(Q)\|_{\Sigma_h^k} = \sum_{i < k} \langle X^{M^*}(\pi^i), W^{M^*}(Q) \rangle^2 = \sum_{i < k} (\mathcal{E}_h(\pi^i, Q))^2 \leq \beta$$

Handwritten notes: $W^{M^}(Q)^T \begin{pmatrix} \epsilon \\ \vdots \\ \epsilon \end{pmatrix} W^{M^*}(Q)$ above the sum; ϵ above the norm; β circled in red.*

- Use Lemma(5) again that $Q^{M^*,*} \in Q^k$, it follows

$$\begin{aligned} \mathbb{E}_{s_1 \sim d_1} [Q_1^k(s_1, \pi_Q(s_1))] &= \sup_{Q \in Q^k} \mathbb{E}_{s_1 \sim d_1} [Q_1(s_1, \pi_Q(s_1))] \\ &\geq \mathbb{E}_{s_1 \sim d_1} [Q_1^{M^*,*}(s_1, \pi_{M^*}(s_1))] \\ &= f^{M^*}(\pi_{M^*}). \end{aligned}$$

Handwritten notes: "def of Q^k " with an arrow pointing to the sup; $Q^{M^,*}$ circled in red; the last two lines underlined in red.*

Proof (Finishing)

Step 1: Suboptimality of π^k for all k

Use the result that Q^k is optimistic above, and set β directly as Lemma(5), we have w.p. $1 - \delta$

$$\begin{aligned}
 f^{M^*}(\pi_{M^*}) - f^{M^*}(\pi^k) &\leq \mathbb{E}_{s_1 \sim d_1} \left[Q_1^k(s_1, \pi_{Q^k}(s_1)) \right] - f^{M^*}(\pi^k) \\
 &= \sum_{h=1}^H \mathbb{E}^{M^*, \pi^k} \left[Q_h^k(s_h, a_h) - r_h - \max_{a \in \mathcal{A}} Q_{h+1}^k(s_{h+1}, a) \right] \\
 &= \sum_{h=1}^H \left\langle X_h^{M^*}(\pi^k), W_h^{M^*}(Q^k) \right\rangle \\
 &\leq \sum_{h=1}^H \left\| X_h^{M^*}(\pi^k) \right\|_{(\lambda I + \Sigma_h^k)^{-1}} \left\| W_h^{M^*}(Q^k) \right\|_{\lambda I + \Sigma_h^k}
 \end{aligned}$$

Handwritten notes:
 - "optimism" with an arrow pointing to the first term.
 - "Bellman decomposition" with a bracket around the first two lines.
 - "Bellman residual decomp" with a bracket around the third line.
 - Red annotations $(\lambda I + \Sigma_h^k)^{-\frac{1}{2}}$ and $\frac{1}{2}$ are placed near the inner product.

Then observe $\left\| W_h^{M^*}(Q^k) \right\|_{\lambda I + \Sigma_h^k} \leq \sqrt{\lambda \left\| W_h^{M^*}(Q^k) \right\|_2^2} + \beta \leq \lambda^{1/2} + \beta^{1/2}$, the general bound follows $(\lambda^{1/2} + \beta^{1/2}) \cdot \sum_{h=1}^H \left\| X_h^{M^*}(\pi^k) \right\|_{(\lambda I + \Sigma_h^k)^{-1}}$. It left to manage 2nd term.

Proof (Finishing) cont'd

A technical lemma helps.

Lemma (7)

For any $\lambda > 0$, as long as $K \geq Hd \log(1 + \lambda^{-1}K/d)$, there exists $k \in [K]$ such that

$$\left\| X_h^{M^*}(\pi^k) \right\|_{(\lambda I + \Sigma_h^k)^{-1}}^2 \lesssim \frac{Hd \log(1 + \lambda^{-1}K/d)}{K} \quad \forall h \in [H].$$

Proof omitted for not being structurally necessary - book page 144. (using elliptic potential lemma, explain...)

- Let $\lambda = \beta$ and let $K \gtrsim Hd \log(1 + n/d)$, the lemma applied. It follows



$$f^{M^*}(\pi_{M^*}) - f^{M^*}(\pi^k) \lesssim H \sqrt{\beta \cdot \frac{Hd \log(1 + \beta^{-1}K/d)}{K}} \lesssim \tilde{O} \left(H^{3/2} \sqrt{\frac{d \log(|\mathcal{Q}|/\delta)}{n}} \right) \lesssim \varepsilon$$

- It's then trivial to see the closure between $\hat{\pi}$ and π^k , which completes the proof.

What's not covered

- Linear Q^*/V^* as example of model with low Bellman rank
- Block MDP (V-type Bellman rank) (if time allows)

References

-  D. J. Foster and A. Rakhlin, “Foundations of reinforcement learning and interactive decision making,” 2023.
-  N. Jiang, “Uiuc cs 542 lecture note,” 2020.