Reinforcement Learning

Mathematical Analysis of Machine Learning Algorithms
(Chapter 18)

Episodic MDP

An episodic Markov decision process (MDP) of length H, denoted by $M = \mathrm{MDP}(\mathcal{X}, \mathcal{A}, P)$, contains a state space $\mathcal{X}$, an action space $\mathcal{A}$, and probability measures $\{P^h(r^h, x^{h+1}|x^h, a^h)\}_{h=1}^H$. At each step $h \in [H] = \{1, \ldots, H\}$, we observe a state $x^h \in \mathcal{X}$ and take action $a^h \in \mathcal{A}$. We then get a reward r^h and go to the next state x^{h+1} with probability $P^h(r^h, x^{h+1}|x^h, a^h)$. We assume that x^1 is drawn from an unknown but fixed distribution.

The goal is to determine action $a^h \in \mathcal{A}$ based on x^h to maximize the reward

$$\sum_{h=1}^{H} [r^h].$$

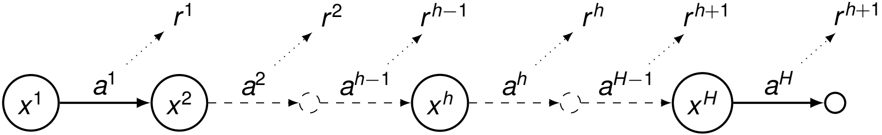

Figure: Episodic Markov decision process

Policy

A random policy π is a set of conditional probability $\pi^h(a^h|x^h)$ that determines the probability of taking action a^h on state x^h at step h. If a policy π is deterministic, then we also write the action a^h it takes at x^h as $\pi^h(x^h) \in \mathcal{A}$.

The policy π interacts with the MDP in an episode as follows: for step $h = 1, \ldots, H$, the player observes x^h, and draws $a^h \sim \pi(a^h|x^h)$; the MDP returns (r^h, x^{h+1}). The reward of the episode is

$$\sum_{h=1}^{H} [r^h].$$

The observations $(x, a, r) = \{(x^h, a^h, r^h)\}_{h=1}^{H}$ is called a trajectory, and each policy π, when interacting with the MDP, defines a distribution over trajectories, which we denote as $(x, a, r) \sim \pi$.

Value of Policy

The value of a policy π is defined as its expected reward:

$$V_\pi = \mathbb{E}_{(x,a,r)\sim\pi} \sum_{h=1}^{H} [r^h].$$

We note that the state x^{H+1} has no significance as the episode ends after taking action a^h at x^h and observe the reward r^h.

Optimal policy value

$$V_* = \sup_\pi V_\pi,$$

with a policy π_* achieving this value referred to as an optimal policy.

Regret

Definition 1

In episodic reinforcement learning (RL), we consider an episodic MDP. The player interacts with the MDP via a repeated game: at each time (episode) t:

- The player chooses a policy π_t based on historic observations.
- The policy interacts with the MDP, and generates a trajectory $(x_t, a_t, r_t) = \{(x_t^h, a_t^h, r_t^h)\}_{h=1}^H \sim \pi_t$.

The regret of episodic reinforcement learning is

$$\sum_{t=1}^{T} [V_* - V_{\pi_t}],$$

where $V_* = \sup_\pi V_\pi$ is the optimal value function.

Example

Example 2 (Contextual Bandits)

Consider the episodic MDP with $H = 1$. We observe $x^1 \in \mathcal{X}$, take action $a^1 \in \mathcal{A}$, and observe reward $r^1 \in \mathbb{R}$. This case is the same as contextual bandits.

Example

Example 3 (Tabular MDP)

In a Tabular MDP, both $\mathcal{X}$ and $\mathcal{A}$ are finite: $|\mathcal{X}| = S$ and $|\mathcal{A}| = A$. It follows that the transition probability at each step h

$$\{P^h(x^{h+1}|x^h, a^h) : h = 1, \ldots, H\}$$

can be expressed using HS^2A numbers. The expected reward $\mathbb{E}[r^h|x^h, a^h]$ can be expressed using HSA numbers.

State and Action Dependent Value Functions

Definition 4

Given any policy π, we can define its value function (also referred to as the Q-function in the literature) starting at a state-action pair (x^h, a^h) at step h as follows:

$$Q_\pi^h(x^h, a^h) = \sum_{h'=h}^{H} \mathbb{E}_{r^{h'} \sim \pi|(x^h, a^h)}[r^{h'}],$$

where $r^{h'} \sim \pi|(x^h, a^h)$ is the reward distribution at step h' conditioned on starting from state action pair (x^h, a^h) at step h. Similarly, we also define

$$V_\pi^h(x^h) = \sum_{h'=h}^{H} \mathbb{E}_{r^{h'} \sim \pi|x^h}[r^{h'}].$$

By convention, we set $V_\pi^{H+1}(x^{H+1}) \equiv 0$.

Property of Value Function

Proposition 5 (Prop 18.7)

We have

$$\begin{aligned} Q_\pi^h(x^h, a^h) =& \mathbb{E}_{r^h, x^{h+1} | x^h, a^h} [r^h + V_\pi^{h+1}(x^{h+1})], \\ V_\pi^h(x^h) =& \mathbb{E}_{a^h \sim \pi^h(\cdot | x^h)} Q_\pi^h(x^h, a^h). \end{aligned}$$

Optimal Value Function

Definition 6

The optimal value functions starting at step h are given by

$$Q_*^h(x^h, a^h) = \sup_\pi Q_\pi^h(x^h, a^h), \qquad V_*^h(x^h) = \sup_\pi V_\pi^h(x^h).$$

We also define the optimal policy value as

$$V_* = \mathbb{E}_{x^1} V_*^1(x^1).$$

Bellman Equation

Theorem 7 (Thm 18.9)

The optimal Q-function Q_* satisfies the Bellman equation:

$$Q_*^h(x^h, a^h) = \mathbb{E}_{r^h, x^{h+1} | x^h, a^h} \left[r^h + V_*^{h+1}(x^{h+1}) \right].$$

The optimal value function satisfies

$$V_*^h(x^h) = \max_{a \in \mathcal{A}} Q_*^h(x^h, a),$$

and the optimal value function can be achieved using a deterministic greedy policy π_* below

$$\pi_*^h(x^h) \in \arg\max_{a \in \mathcal{A}} Q_*^h(x^h, a).$$

Bellman Error

Definition 8

We say f is a candidate Q-function if
$f = \{f^h(x^h, a^h) : \mathcal{X} \times \mathcal{A} \to \mathbb{R} : h \in [H+1]\}$, with $f^{H+1}(\cdot) = 0$. Define

$$f^h(x^h) = \arg\max_{a \in \mathcal{A}} f^h(x^h, a),$$

and define its greedy policy π_f as a deterministic policy that satisfies

$$\pi_f^h(x^h) \in \arg\max_{a \in \mathcal{A}} f^h(x^h, a).$$

Given an MDP M, we also define the Bellman operator of f as

$$(\mathcal{T}^h f)(x^h, a^h) = \mathbb{E}_{r^h, x^{h+1} | x^h, a^h}[r^h + f^{h+1}(x^{h+1})],$$

and its Bellman error as

$$\mathcal{E}^h(f, x^h, a^h) = f^h(x^h, a^h) - (\mathcal{T}^h f)(x^h, a^h),$$

where the conditional expectation is with respect to the MDP M.

Value Decomposition

We note

$$\mathcal{E}^h(Q_*, x^h, a^h) = 0, \quad \forall h \in [H].$$

The following result shows that the reverse is also true.

Theorem 9 (Thm 18.11)

Consider any candidate value function $f = \{f^h(x^h, a^h) : \mathcal{X} \times \mathcal{A} \to \mathbb{R}\}$*, with* $f^{H+1}(\cdot) = 0$*. Let* π_f *be its greedy policy. Then*

$$[f^1(x^1) - V^1_{\pi_f}(x^1)] = \mathbb{E}_{(x,a,r)\sim\pi_f|x^1} \sum_{h=1}^{H} \mathcal{E}^h(f, x^h, a^h).$$

Proof of Theorem 9 (I/II)

We prove the following statement by induction from $h = H$ to $h = 1$.

$$[f^h(x^h) - V^h_{\pi_f}(x^h)] = \mathbb{E}_{\{(x^{h'},a^{h'},r^{h'})\}_{h'=h}^{H} \sim \pi_f | x^h} \sum_{h'=h}^{H} \mathcal{E}^{h'}(f, x^{h'}, a^{h'}). \quad (1)$$

When $h = H$, we have $a^H = \pi_f^H(x^H)$ and

$$\mathcal{E}^H(f, x^H, a^H) = f^H(x^H, a^H) - \mathbb{E}_{r^H|x^H,a^H}[r^H] = f^H(x^H) - V_\pi^H(x^H).$$

Therefore (1) holds.

Proof of Theorem 9 (II/II)

Assume that the equation holds at $h+1$ for some $1 \le h \le H-1$. Then at h, we have

$$
\begin{aligned}
&\mathbb{E}_{\{(x^{h'},a^{h'},r^{h'})\}_{h'=h}^{H} \sim \pi_f | x^h} \sum_{h'=h}^{H} \mathcal{E}^{h'}(f, x^{h'}, a^{h'}) \\
=&\mathbb{E}_{x^{h+1}, r^h, a^h \sim \pi_f | x^h}[\mathcal{E}^h(f, x^h, a^h) + f^{h+1}(x^{h+1}) - V_{\pi_f}^{h+1}(x^{h+1})] \\
=&\mathbb{E}_{x^{h+1}, r^h, a^h \sim \pi_f | x^h}[f^h(x^h, a^h) - r^h - V_{\pi_f}^{h+1}(x^{h+1})] \\
=&\mathbb{E}_{a^h \sim \pi_f | x^h}[f^h(x^h, a^h) - V_{\pi_f}^{h}(x^h)] \\
=&[f^h(x^h) - V_{\pi_f}^{h}(x^h)].
\end{aligned}
$$

The first equation used the induction hypothesis. The second equation used the definition of Bellman error. The third equation used Proposition 5. The last equation used $a^h = \pi_f(x^h)$ and thus by definition, $f^h(x^h, a^h) = f^h(x^h)$.

Realizable Assumption

Assumption 10 (Asm 18.12)

Given a candidate value function class $\mathcal{F}$ *of functions* $f = \{f^h(x^h, a^h) : \mathcal{X} \times \mathcal{A} \to \mathbb{R}\}$*, with* $f^{H+1}(\cdot) = 0$*. We assume that (realizable assumption)*

$$Q_* = f_* \in \mathcal{F}.$$

Moreover, we assume that $f^1(x^1) \in [0, 1]$ *and* $r^h + f^{h+1}(x^{h+1}) \in [0, 1]$ *(*$h \geq 1$*).*

Completeness Assumption

Definition 11 (Bellman Completeness)

A candidate value function class $\mathcal{F}$ is complete with respect to another candidate value function class $\mathcal{G}$ if for any $h \in [H]$, $f \in \mathcal{F}$, there exists $g \in \mathcal{G}$ so that for all $h \in [H]$:

$$g^h(x^h, a^h) = (\mathcal{T}^h f)(x^h, a^h) = \mathbb{E}_{r^h, x^{h+1} | x^h, a^h} \left[r^h + f^{h+1}(x^{h+1}) \right].$$

We say $\mathcal{F}$ is complete if $\mathcal{F}$ is complete with respect to itself.

Linear MDP

Definition 12 (Linear MDP, Def 18.15)

Let $\mathcal{H} = \{\mathcal{H}^h\}$ be a sequence of vector spaces with inner products $\langle \cdot, \cdot \rangle$. An MDP $M = \mathrm{MDP}(\mathcal{X}, \mathcal{A}, P)$ is a linear MDP with feature maps $\phi = \{\phi^h(x^h, a^h) : \mathcal{X} \times \mathcal{A} \to \mathcal{H}^h\}_{h=1}^H$ if for all $h \in [H]$, there exist a map $\nu^h(x^{h+1}) : \mathcal{X} \to \mathcal{H}^h$ and $\theta^h \in \mathcal{H}^h$, such that

$$dP^h(x^{h+1}|x^h, a^h) = \langle \nu^h(x^{h+1}), \phi^h(x^h, a^h) \rangle d\mu^{h+1}(x^{h+1}),$$
$$\mathbb{E}[r^h|x^h, a^h] = \langle \theta^h, \phi^h(x^h, a^h) \rangle.$$

Here $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathcal{H}^h$ for different h, and the conditional probability measure $dP^h(\cdot|x^h, a^h)$ is absolute continuous with respect to a measure $d\mu^{h+1}(\cdot)$ with density $\langle \nu^h(x^{h+1}), \phi^h(x^h, a^h) \rangle$. In general, we assume that $\nu^h(\cdot)$ and θ^h are unknown.

We assume $\phi(\cdot)$ is either known or unknown.

Example

Example 13 (Tabular MDP)

In a tabular MDP, we assume that $|\mathcal{A}| = A$ and $|\mathcal{X}| = S$. Let $d = AS$, and we can encode the space of $\mathcal{X} \times \mathcal{A}$ into a d-dimensional vector with components indexed by (x, a). Let $\phi^h(x, a) = e_{(x,a)}$ and let $\nu^h(x^{h+1})$ be a d dimensional vector so that its (x, a) component is $P^h(x^{h+1}|x^h = x, a^h = a)$. Similarly, we can take θ^h as a d dimensional vector so that its (x, a) component is $\mathbb{E}[r^h|x^h = x, a^h = a]$. Therefore tabular MDP is linear MDP with $d = AS$.

Example

Example 14 (Low-Rank MDP)

For a low-rank MDP, we assume that the transition probability matrix can be decomposed as

$$P^h(x^{h+1}|x^h, a^h) = \sum_{j=1}^{d} P^h(x^{h+1}|z = j)P^h(z = j|x^h, a^h).$$

In this case we can set $\phi^h(x^h, a^h) = [P^h(z = j|x^h, a^h)]_{j=1}^d$, and $\nu^h(x^{h+1}) = [P^h(x^{h+1}|z = j)]_{j=1}^d$. Therefore a low-rank MDP is a linear MDPs with rank as dimension.

Property of Linear MDP

Proposition 15 (Prop 18.18)

In a linear MDP with feature map $\phi^h(x^h, a^h)$ on vector spaces $\mathcal{H}^h$ ($h \in [H]$). Consider the linear candidate Q function class

$$\mathcal{F} = \left\{ \langle w^h, \phi^h(x^h, a^h) \rangle : w^h \in \mathcal{H}^h, h \in [H] \right\}.$$

Any function $g^{h+1}(x^{h+1})$ on $\mathcal{X}$ satisfies

$$(\mathcal{T}^h g^{h+1})(x^h, a^h) \in \mathcal{F}.$$

It implies that $\mathcal{F}$ is complete, and $Q_* \in \mathcal{F}$. Moreover, $\forall f \in \mathcal{F}$,

$$\mathcal{E}^h(f, x^h, a^h) \in \mathcal{F}.$$

Proof of Proposition 15

Let

$$u_g^h = \int g^{h+1}(x^{h+1})\nu^h(x^{h+1})d\mu^{h+1}(x^{h+1}).$$

We have

$$\begin{aligned}\mathbb{E}_{x^{h+1}|x^h,a^h} g^{h+1}(x^{h+1}) =& \int g^{h+1}(x^{h+1})\langle \nu^h(x^{h+1}), \phi^h(x^h,a^h)\rangle d\mu^{h+1}(x^{h+1}) \\ =& \langle u_g^h, \phi^h(x^h,a^h)\rangle.\end{aligned}$$

This implies that

$$(\mathcal{T}^h g)(x^h,a^h) = \langle \theta^h + u_g^h, \phi^h(x^h,a^h)\rangle \in \mathcal{F}.$$

Since $Q_*^h(x^h,a^h) = (\mathcal{T}^h Q_*)(x^h,a^h)$, we know $Q_*^h(x^h,a^h) \in \mathcal{F}$. Similarly, since $(\mathcal{T}^h f)(x^h,a^h) \in \mathcal{F}$, we know that $f \in \mathcal{F}$ implies

$$\mathcal{E}^h(f,x^h,a^h) = f^h(x^h,a^h) - (\mathcal{T}^h f)(x^h,a^h) \in \mathcal{F}.$$

This proves the desired result.

Estimating Bellman Error

Consider

$$(f^h(x^h, a^h) - r^h - f^{h+1}(x^{h+1}))^2. \tag{2}$$

By taking conditional expectation with respect to (x^h, a^h), we obtain

$$\begin{aligned}&\mathbb{E}_{r^h, x^{h+1}|x^h, a^h}(f^h(x^h, a^h) - r^h - f^{h+1}(x^{h+1}))^2 \\ =&\mathcal{E}^h(f, x^h, a^h)^2 + \mathbb{E}_{r^h, x^{h+1}|x^h, a^h}\Big(\underbrace{r^h + f^{h+1}(x^{h+1}) - (\mathcal{T}^h f)(x^h, a^h)}_{f\text{-dependent zero-mean noise}}\Big)^2.\end{aligned}$$

Since noise variance depends on f, if we use (2) to estimate f, we will favor f with smaller noise variance, which may not have zero Bellman error.

The Role of Completeness in Bellman Error Estimation

If $\mathcal{F}$ is complete with respect to $\mathcal{G}$, then we may use the solution of

$$\min_{g^h \in \mathcal{G}^h} \sum_{s=1}^{t} (g^h(x_s^h, a_s^h) - r_s^h - f^{h+1}(x_s^{h+1}))^2$$

to estimate $(\mathcal{T}^h f)(x^h, a^h)$, which can be used to cancel the f dependent variance term in (2).

This motivates the following loss function

$$L^h(f, g, x^h, a^h, r^h, x^{h+1}) = \Big[(f^h(x^h, a^h) - r^h - f^{h+1}(x^{h+1}))^2 - (g^h(x^h, a^h) - r^h - f^{h+1}(x^{h+1}))^2\Big]. \quad (3)$$

We have

$$\sup_{g \in \mathcal{G}} \sum_{h=1}^{H} \sum_{s=1}^{t} L^h(f, g, x_s^h, a_s^h, r_s^h, x_s^{h+1}) \approx \sum_{h=1}^{H} \sum_{s=1}^{t} \mathcal{E}^h(f, x_s^h, a_s^h)^2.$$

Property of Minimax Bellman Error Estimator

Theorem 16 (Thm 18.14)

Assume that assumption 10 holds, $\mathcal{F}$ is complete with respect to $\mathcal{G}$, and $g^h(\cdot) \in [0, 1]$ for all $g \in \mathcal{G}$. Consider (3)*, and let*

$$\mathcal{F}_t = \left\{ f \in \mathcal{F} : \sup_{g \in \mathcal{G}} \sum_{h=1}^{H} \sum_{s=1}^{t} L^h(f, g, x_s^h, a_s^h, r_s^h, x_s^{h+1}) \leq \beta_t^2 \right\},$$

where

$$\beta_t^2 \geq 4\epsilon t(4+\epsilon)H + 2\ln\left(16 M(\epsilon, \mathcal{F}, \|\cdot\|_\infty)^2 M(\epsilon, \mathcal{G}, \|\cdot\|_\infty)/\delta^2\right),$$

with $M(\cdot)$ denotes the $\|\cdot\|_\infty$ packing number, and $\|f\|_\infty = \sup_{h,x,a} |f^h(x, a)|$. Then with probability at least $1-\delta$, for all $t \leq n$: $Q_* \in \mathcal{F}_t$ and for all $f \in \mathcal{F}_t$:

$$\sum_{s=1}^{t} \sum_{h=1}^{H} \mathcal{E}^h(f, x_s^h, a_s^h)^2 \leq 4\beta_t^2.$$

UCB Algorithm

Algorithm 1: Bellman Error UCB Algorithm

Input: λ, T, $\mathcal{F}$, $\mathcal{G}$

Let $\mathcal{F}_0 = \{f_0\}$

Let $\beta_0 = 0$

for $t = 1, 2, \ldots, T$ **do**

Observe x_t^1

Let $f_t \in \arg\max_{f \in \mathcal{F}_{t-1}} f(x_t^1)$.

Let $\pi_t = \pi_{f_t}$

Play policy π_t and observe trajectory (x_t, a_t, r_t)

Let

$$\mathcal{F}_t = \left\{ f \in \mathcal{F} : \sup_{g \in \mathcal{G}} \sum_{h=1}^{H} \sum_{s=1}^{t} L^h(f, g, x_s^h, a_s^h, r_s^h, x_s^{h+1}) \le \beta_t^2 \right\}$$

with appropriately chosen β_t, where $L^h(\cdot)$ is defined in (3).

return randomly chosen π_t from $t = 1$ to $t = T$

Analysis of Algorithm 1: Eluder Coefficient

Definition 17 (Q-type Bellman Eluder Coefficient, Def 18.19)

Given a candidate Q function class $\mathcal{F}$, its Q-type Bellman eluder coefficient $\mathrm{EC}_Q(\epsilon, \mathcal{F}, T)$ is the smallest number d so that for any filtered sequence $\{f_t, (x_t, r_t, a_t) \sim \pi_{f_t}\}_{t=1}^{T}$:

$$\mathbb{E} \sum_{t=2}^{T} \sum_{h=1}^{H} \mathcal{E}^h(f_t, x_t^h, a_t^h) \leq \sqrt{d \, \mathbb{E} \sum_{h=1}^{H} \sum_{t=2}^{T} \left(\epsilon + \sum_{s=1}^{t-1} \mathcal{E}^h(f_t, x_s^h, a_s^h)^2 \right)}.$$

Eluder Coefficient for Linear MDP

Proposition 18 (Simplification of Prop 18.20)

Assume that a linear MDP has (possibly unknown) d^h dimensional feature maps $\phi^h(x^h, a^h)$ for each h.
Assume also that the candidate Q-function class $\mathcal{F}$ can be embedded into the linear function space

$$\mathcal{F} \subset \{\langle w^h, \phi^h(x^h, a^h)\rangle : w^h \in \mathcal{H}^h\},$$

and there exists $B > 0$ such that $\|\mathcal{E}^h(f, \cdot, \cdot)\|_{\mathcal{H}^h} \leq B$.
Assume that $|\mathcal{E}^h(f, x^h, a^h)| \in [0, 1]$, then

$$\mathrm{EC}_Q(1, \mathcal{F}, T) \leq 2 \sum_{h=1}^{H} d^h \ln(1 + T(BB')^2),$$

where $B' = \sup_h \sup_{x^h, a^h} \|\phi^h(x^h, a^h)\|_{\mathcal{H}^h}$.

Regret Bound

Theorem 19 (Thm 18.21)

Assume that Assumption 10 holds, $\mathcal{F}$ is complete with respect to $\mathcal{G}$, and $g^h(\cdot) \in [0, 1]$ for all $g \in \mathcal{G}$. Assume also that β_t is chosen in Algorithm 1 according to

$$\beta_t^2 \geq \inf_{\epsilon > 0} \left[4\epsilon t(4 + \epsilon)H + 2 \ln \left(16 M(\epsilon, \mathcal{F}, \| \cdot \|_\infty)^2 M(\epsilon, \mathcal{G}, \| \cdot \|_\infty) / \delta^2 \right) \right],$$

with $M(\cdot)$ denoting the $\| \cdot \|_\infty$ packing number, and $\|f\|_\infty = \sup_{h,x,a} |f^h(x, a)|$. Then

$$\begin{aligned} &\mathbb{E} \sum_{t=2}^{T} [V_*^1(x_t^1) - V_{\pi_t}^1(x_t^1)] \\ \leq &\delta T + \sqrt{\mathrm{EC}_Q(\epsilon, \mathcal{F}, T) \left(\epsilon H T + \delta H T^2 + 4 \sum_{t=2}^{T} \beta_{t-1}^2 \right)}. \end{aligned}$$

Proof of Theorem 19 (I/II)

For $t \geq 2$, we have

$$\begin{aligned}
& V_*^1(x_t^1) - V_{\pi_t}^1(x_t^1) \\
=& V_*^1(x_t) - f_t(x_t^1) + f_t(x_t^1) - V_{\pi_t}^1(x_t^1) \\
\leq& \mathbb{1}(Q_* \notin \mathcal{F}_{t-1}) + [f_t(x_t^1) - V_{\pi_t}^1(x_t^1)] \\
=& \mathbb{1}(Q_* \notin \mathcal{F}_{t-1}) + \mathbb{E}_{(x_t, a_t, r_t) \sim \pi_t | x_t^1} \sum_{h=1}^{H} \mathcal{E}^h(f_t, x_t^h, a_t^h).
\end{aligned}$$

The inequality used the fact that if $Q_* \in \mathcal{F}_{t-1}$, then $f_t(x_t^1) = \max_{f \in \mathcal{F}_{t-1}} f(x_t^1) \geq V_*^1(x_t^1)$, and if $Q_* \notin \mathcal{F}_{t-1}$, $V_*^1(x_t) - f_t(x_t^1) \leq 1$. The last equation used Theorem 9. Theorem 16 implies that $\Pr(Q_* \in \mathcal{F}_{t-1}) \geq 1 - \delta$. We thus have

$$\mathbb{E}[V_*^1(x_t^1) - V_{\pi_t}^1(x_t^1)] \leq \delta + \mathbb{E} \sum_{h=1}^{H} \mathcal{E}^h(f_t, x_t^h, a_t^h).$$

Proof of Theorem 19 (II/II)

We can now obtain

$$\begin{aligned}
&\mathbb{E} \sum_{t=2}^{T} [V_*^1(x_t^1) - V_{\pi_t}^1(x_t^1)] \\
\leq &\mathbb{E} \sum_{t=2}^{T} \sum_{h=1}^{H} \mathcal{E}^h(f_t, x_t^h, a_t^h) + \delta T \\
\leq &\delta T + \sqrt{\mathrm{EC}_Q(\epsilon, \mathcal{F}, T) \mathbb{E} \sum_{t=2}^{T} \sum_{h=1}^{H} \left(\epsilon + \sum_{s=1}^{t-1} \mathcal{E}^h(f_t, x_s^h, a_s^h)^2 \right)} \\
\leq &\delta T + \sqrt{\mathrm{EC}_Q(\epsilon, \mathcal{F}, T) \left(\epsilon H T + \delta H T^2 + 4 \sum_{t=2}^{T} \beta_{t-1}^2 \right)}.
\end{aligned}$$

The second inequality used Definition 17. The last inequality used the fact that for each t, Theorem 16 holds with probability $1 - \delta$, and otherwise, $\mathcal{E}^h(f_t, x_s^h, a_s^h)^2 \leq 1$.

Interpretation of Theorem 19: Linear MDP

Consider the d dimensional linear MDP with bounded $\mathcal{F}$ and $\mathcal{G}$. Assume that the model coefficients at different step h are different, then the entropy can be bounded (ignoring log factors) as

$$\tilde{O}(H \ln(M_{\mathcal{F}} M_{\mathcal{G}})) = \tilde{O}(Hd),$$

and hence with $\epsilon = \delta = O(1/T^2)$, we have

$$\beta_t^2 = \tilde{O}(H \ln(M_{\mathcal{F}} M_{\mathcal{G}})) = \tilde{O}(Hd).$$

Since $\mathrm{EC}_Q(\epsilon, \mathcal{F}, T) = \tilde{O}(dH)$, we obtain the following.

Regret Bound from Theorem 19

We have the following regret bound for Algorithm 1

$$\mathbb{E}\,\mathrm{REG}_T = \tilde{O}\left(H\sqrt{dT \ln(M_{\mathcal{F}} M_{\mathcal{G}})}\right) = \tilde{O}\left(Hd\sqrt{T}\right). \tag{4}$$

Least Squares Value Iteration

It was shown in Theorem 19 that the UCB method in Algorithm 1 can handle linear MDP with Q-type Bellman eluder coefficient. However, it requires solving a minimax formulation with global optimism, which may be difficult computationally. In fact, there is no practically effective implementation of the method.

Next, we show that a computationally more efficient procedure, referred to as Least Squares Value Iteration (LSVI), or Fitted Q-learning, can be used to solve RL. This procedure is closely related to the Q-learning method used by practitioners.

Assumption for LSVI Algorithm

Assumption 20 (Completeness, Asm 18.22)

Assume that the Q function class $\mathcal{F}$ can be factored as the product of H function classes:

$$\mathcal{F} = \prod_{h=1}^{H} \mathcal{F}^h, \quad \mathcal{F}^h = \{\langle w^h, \phi^h(x^h, a^h)\rangle, w^h \in \mathcal{H}^h\},$$

so that for all $g^{h+1}(x^{h+1}) \in [0, 1]$:

$$(\mathcal{T}^h g^{h+1})(x^h, a^h) \in \mathcal{F}^h. \tag{5}$$

Assumption for LSVI Algorithm

Assumption 21 (Bonus Function, Asm 18.22)

In Assumption 20, assume further for any $\epsilon > 0$, there exists a function class $\mathcal{B}^h(\epsilon)$ so that for any sequence $\{(x_t^h, a_t^h, \hat{f}_t^h) \in \mathcal{X} \times \mathcal{A} \times \mathcal{F}^h : t = 1, \ldots, T\}$, we can construct a sequence of non-negative bonus functions $b_t^h(\cdot) \in \mathcal{B}^h(\epsilon)$ (each $\hat{f}_t^h$ and b_t^h only depend on the historic observations up to $t-1$) such that

$$b_t^h(x^h, a^h)^2 \geq \sup_{f^h \in \mathcal{F}^h} \frac{|f^h(x^h, a^h) - \hat{f}_t^h(x^h, a^h)|^2}{\epsilon + \sum_{s=1}^{t-1} |f^h(x_s^h, a_s^h) - \hat{f}_t^h(x_s^h, a_s^h)|^2}, \tag{6}$$

and the bonus function satisfies the following uniform eluder condition*:*

$$\sup_{\{(x_t^h, a_t^h)\}} \sum_{t=1}^{T} \min(1, b_t^h(x_t^h, a_t^h)^2) \leq \dim(T, \mathcal{B}^h(\epsilon)).$$

Example 18.23: Linear MDP (I/II)

Consider a linear MDP in Definition 12, such that

$$\|\theta^h\|_{\mathcal{H}^h} + \int \|\nu^h(x^{h+1})\|_{\mathcal{H}^h} \, |d\mu^{h+1}(x^{h+1})| \leq B^h.$$

If $\mathcal{F}^h$ is any function class that contains

$$\tilde{\mathcal{F}}^h = \{\langle w^h, \phi^h(x^h, a^h)\rangle : \|w^h\|_{\mathcal{H}^h} \leq B^h\},$$

then the proof of Proposition 15 implies that (5) holds.
Note that if $r^h \in [0, 1]$, then $(\mathcal{T}^h g^{h+1})(x^h, a^h) \in [0, 2]$. Therefore at any time step t, we may consider a subset of $\mathcal{F}^h$ that satisfies the range constraint on historic observations, and in the mean time, impose the same range constraints in $\tilde{\mathcal{F}}^h$ as

$$\tilde{\mathcal{F}}^h = \Big\{\langle w^h, \phi^h(x^h, a^h)\rangle : \|w^h\|_{\mathcal{H}^h} \leq B^h, \\ \langle w^h, \phi^h(x^h_s, a^h_s)\rangle \in [0, 2] \, \forall s \in [t-1]\Big\}.$$

Example 18.23: Linear MDP (II/II)

If moreover, each $f^h(x^h, a^h) \in \mathcal{F}^h$ can be written as $\langle \tilde{w}^h(f^h), \tilde{\phi}^h(x^h, a^h) \rangle$ so that $\|\tilde{w}^h(f^h) - \tilde{w}^h(\tilde{f}^h)\|_2 \leq \tilde{B}^h$ (here we assume that $\tilde{\phi}^h$ may or may not be the same as ϕ^h), then we can take

$$b_t^h(x^h, a^h) = \|\tilde{\phi}^h(x^h, a^h)\|_{(\Sigma_t^h)^{-1}}, \tag{7}$$

$$\Sigma_t^h = \frac{\epsilon}{(\tilde{B}^h)^2} I + \sum_{s=1}^{t-1} \tilde{\phi}^h(x^h, a^h) \tilde{\phi}^h(x^h, a^h)^\top,$$

so that (6) holds. By using Lemma 13.9, we have

$$\begin{aligned} \sum_{t=1}^{T} \min\left(1, \|\tilde{\phi}^h(x_t^h, a_t^h)\|^2_{(\Sigma_t^h)^{-1}}\right) \leq & \sum_{t=1}^{T} \frac{2\|\tilde{\phi}^h(x_t^h, a_t^h)\|^2_{(\Sigma_t^h)^{-1}}}{1 + \|\tilde{\phi}^h(x_t^h, a_t^h)\|^2_{(\Sigma_t^h)^{-1}}} \\ \leq & \ln \left| ((\tilde{B}^h)^2/\epsilon)\Sigma_t^h \right|. \end{aligned}$$

Using Proposition 15.8, we can set $\dim(T, \mathcal{B}^h(\epsilon)) = \mathrm{entro}\big(\epsilon/((\tilde{B}^h)^2 T), \tilde{\phi}^h(\cdot)\big)$. For d dimensional problem, $\dim(T, \mathcal{B}^h(\epsilon)) = \tilde{O}(d)$.

Linear Least Squares Value Iteration

Algorithm 2: Least Squares Value Iteration with UCB (LSVI-UCB)

Input: $\epsilon > 0$, T, $\{\mathcal{F}^h\}$, $\{\mathcal{B}^h(\epsilon)\}$

for $t = 1, 2, \ldots, T$ **do**

  Let $f_t^{H+1} = 0$

  for $h = H, H-1, \ldots, 1$ **do**

    Let $y_s^h = r_s^h + f_t^{h+1}(x_s^{h+1})$, where $f_t^{h+1}(x_s^{h+1}) = \max_a f_t^{h+1}(x_s^{h+1}, a)$

    Let

$$\hat{f}_t^h = \arg\min_{f^h \in \mathcal{F}^h} \sum_{s=1}^{t-1} (f^h(x_s^h, a_s^h) - y_s^h)^2.$$

    Find $\beta_t^h > 0$ and bonus function $b_t^h(\cdot)$ that satisfies (6)

    Let $f_t^h(x^h, a^h) = \min(1, \max(0, \hat{f}_t^h(x^h, a^h) + \beta_t^h b_t^h(x^h, a^h)))$

  Let π_t be the greedy policy of f_t^h for each step $h \in [H]$

  Play policy π_t and observe trajectory (x_t, a_t, r_t)

return randomly chosen π_t from $t = 1$ to $t = T$

Analysis of LSVI-UCB: Key Lemma

Lemma 22 (Lem 18.24)

Consider Algorithm 2 under Assumption 18.22. Assume also that $Q_*^h \in \mathcal{F}^h$, $Q_*^h \in [0,1]$, $r^h \in [0,1]$, $f^h \in [0,2]$ *for* $h \in [H]$ *and* $f^h \in \mathcal{F}^h$*. Given any* $t > 0$, *let* $\beta_t^{H+1} = \beta^{H+1}(\epsilon, \delta) = 0$, *and for* $h = H, H-1, \ldots, 1$:

$$\beta_t^h = \beta^h(\epsilon, \delta) \geq 4(1 + \beta^{h+1})\frac{\epsilon}{\sqrt{T}} + \sqrt{\epsilon} + \sqrt{24(1 + \beta^{h+1}(\delta))\epsilon + 12 \ln \frac{2H\, M_T^h(\epsilon)}{\delta}},$$

where (with $\|f\|_\infty = \sup_{x,a,h} f^h(x,a)$*)*

$$M_T^h(\epsilon) = M(\epsilon/T, \mathcal{F}^h, \|\cdot\|_\infty) M(\epsilon/T, \mathcal{F}^{h+1}, \|\cdot\|_\infty) M(\epsilon/T, \mathcal{B}^{h+1}(\epsilon), \|\cdot\|_\infty).$$

Then with probability at least $1 - \delta$, *for all* $h \in [H]$, *and* $(x^h, a^h) \in \mathcal{X} \times \mathcal{A}$:

$$Q_*^h(x^h, a^h) \leq f_t^h(x^h, a^h),$$
$$|f_t^h(x^h, a^h) - (\mathcal{T}^h f_t^{h+1})(x^h, a^h)| \leq 2\beta^h(\epsilon, \delta) b^h(x^h, a^h).$$

Regret Bound for LSVI-UCB

Theorem 23 (Thm 18.25)

Consider Algorithm 2, and assume that all conditions of Lemma 22 hold. Then

$$\mathbb{E}\sum_{t=1}^{T}[V_*^1(x_t^1) - V_{\pi_t}^1(x_t^1)] \leq \delta T + 2\sqrt{dHT\sum_{h=1}^{H}\beta^h(\epsilon,\delta)^2} + 2Hd,$$

where $d = H^{-1}\sum_{h=1}^{H}\dim(T, \mathcal{B}^h(\epsilon))$.

Proof of Theorem 23 (I/II)

From Lemma 22, we know that for each t, with probability at least $1-\delta$ over the observations $\{(x_s, a_s, r_s) : s = 1, \ldots, t-1\}$, the two inequalities of the lemma hold (which we denote as event E_t). It implies that under event E_t, f_t^h satisfies the following inequalities for all $h \in [H]$:

$$\mathbb{E}_{x_t^1} V_*^1(x_t^1) \leq \mathbb{E}_{x_t^1} f_t^1(x_t^1), \tag{8}$$

$$\mathbb{E}_{x_t^h, a_t^h} |\mathcal{E}^h(f_t, x_t^h, a_t^h)| \leq 2\mathbb{E}_{x_t^h, a_t^h} \beta^h(\epsilon, \delta) b^h(x_t^h, a_t^h). \tag{9}$$

Proof of Theorem 23 (II/II)

We thus obtain

$$
\begin{aligned}
\mathbb{E}\sum_{t=1}^{T}[V_*^1(x_t^1) - V_{\pi_t}^1(x_t^1)] \leq \delta T + \mathbb{E}\sum_{t=1}^{T}[f_t^1(x_t^1) - V_{\pi_t}^1(x_t^1)]\mathbb{1}(E_t) \\
= & \delta T + \sum_{t=1}^{T} \mathbb{E}\sum_{h=1}^{H} \mathcal{E}^h(f_t, x_t^h, a_t^h)\mathbb{1}(E_t) \\
\leq & \delta T + 2\sum_{t=1}^{T} \mathbb{E} \sum_{h=1}^{H} \left[\beta^h(\epsilon,\delta)\min(1, b^h(x_t^h, a_t^h)) + \min(1, b^h(x_t^h, a_t^h))^2\right] \\
\leq & \delta T + 2\sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H}\beta^h(\epsilon,\delta)^2}\sqrt{\mathbb{E}\sum_{t=1}^{T}\sum_{h=1}^{H}\min(1, b^h(x_t^h, a_t^h))^2} \\
& + 2\mathbb{E}\sum_{t=1}^{T}\sum_{h=1}^{H}\min(1, b^h(x_t^h, a_t^h))^2 \\
\leq & \delta T + 2\sqrt{T\sum_{h=1}^{H}\beta^h(\epsilon,\delta)^2}\sqrt{\sum_{h=1}^{H}\dim(T, \mathcal{B}^h(\epsilon))} + 2\sum_{h=1}^{H}\dim(T, \mathcal{B}^h(\epsilon)).
\end{aligned}
$$

Interpretation of Theorem 23 : Linear MDP

Consider linear MDP with known d dimensional $\phi^h(\cdot) = \tilde{\phi}^h(\cdot)$.

- We have $\ln N(\epsilon/T, \mathcal{F}^h, \|\cdot\|_\infty) = \tilde{O}(d)$.
- Since the bonus function of (7) can be regarded as a function class with the $d \times d$ matrix Σ_t^h as its parameter, Theorem 5.3 implies $\ln N(\epsilon/T, \mathcal{B}^{h+1}(\epsilon), \|\cdot\|_\infty) = \tilde{O}(d^2)$.
- We have $\dim(T, \mathcal{B}^h(\epsilon)) = \tilde{O}(d)$ from Example 18.23 and Proposition 15.8. We can set $\beta^h = \tilde{O}(d^2)$.

Regret Bound from Theorem 23

For Algorithm 2, we have

$$\mathbb{E}\,\mathrm{REG}_T = \tilde{O}(Hd^{3/2}\sqrt{T}).$$

The bound is inferior by a factor of $\sqrt{d}$ compared to (4), due to the $\tilde{O}(d^2)$ entropy number of the bonus function class $\mathcal{B}^{h+1}(\epsilon)$.

Model Based RL

Definition 24 (Def 18.35)

In a model-based RL problem, we are given an MDP model class $\mathcal{M}$. Each $M \in \mathcal{M}$ includes explicit transition probability

$$P_M^h(x^{h+1}|x^h, a^h),$$

and expected reward

$$R_M^h(x^h, a^h) = \mathbb{E}_M\left[r^h|x^h, a^h\right],$$

where we use $\mathbb{E}_M[\cdot]$ to denote the expectation with respect to model M's transition dynamics P_M.
We use $f_M = \{f_M^h(x^h, a^h)\}_{h=1}^H$ to denote the Q function of model M, and use $\pi_M = \pi_{f_M}$ to denote the corresponding optimal policy under model M.

Example: Linear Mixture MDP

A simple example of model-based reinforcement learning problem is linear mixture MDP (also see Definition 18.48).

Example 25 (Mixture of Known MDPs, Expl 18.50)

Consider d base MDPs $M_1, \ldots, M_d$, where each MDP M_j corresponds to a transition distribution $P^h_{M_j}(x^{h+1}|x^h, a^h)$ and an expected reward $R^h_{M_j}(x^h, a^h)$. Consider a model family $\mathcal{M}$, where $M \in \mathcal{M}$ is represented by $w_1, \ldots, w_d \geq 0$ and $\sum_{j=1}^d w_j = 1$. Then we can express

$$P^h_M(x^{h+1}|x^h, a^h) = \sum_{j=1}^{d} w_j P^h_{M_j}(x^{h+1}|x^h, a^h).$$

One can similarly define $R^h_M(x^h, a^h) = \sum_{j=1}^d w_j R^h_{M_j}(x^h, a^h)$.

Generic Model-Based Algorithm

Algorithm 3: *Q*-type Model-Based Posterior Sampling Algorithm

Input: λ, η, $\tilde{\eta}$, T, p_0, $\mathcal{M}$

for $t = 1, 2, \ldots, T$ **do**

Observe x_t^1

Draw

$$M_t \sim p_t(M|x_t^1, S_{t-1})$$

according to $p_t(M|x_1^t, S_{t-1})$ defined as

$$p_t(M|x_1^t, S_{t-1}) \propto p_0(M) \exp\left(\lambda \sum_{s=1}^{t-1} f_M(x_s^1) + \sum_{h=1}^{H} \sum_{s=1}^{t-1} L_s^h(M)\right),$$

$$L_s^h(M) = -\tilde{\eta}(R_M^h(x_s^h, a_s^h) - r_s^h)^2 + \eta \ln P_M^h(x_s^{h+1} \mid x_s^h, a_s^h).$$

Let $\pi_t = \pi_{M_t}$

Play policy π_t and observe trajectory (x_t, a_t, r_t)

Analysis of Mixture of Known MDPs

The analysis of Algorithm 3 can be found in Theorem 18.47.

For Mixture of Known MDPs, we can obtain the following result.

Regret Bound from Theorem 18.47

If we apply Algorithm 3 to Example 25 with appropriate parameter choices, then

$$\mathbb{E}\,\mathrm{REG}_T = \tilde{O}(dH\sqrt{T}).$$

This result is similar to that of linear MDP.

Summary (Chapter 18)

- Episodic Reinforcement Learning
- Policy and Value Function
- Bellman Equation
- Realizability and Completeness
- Linear MDP
- UCB Algorithm for (Model Free) Episodic RL
- LSVI Algorithm for (Model Free) Episodic RL
- Model Based RL