Reinforcement Learning: Basics

Chaowen Zheng

Bandit Algorithm Reading Groups

May 10, 2024

Outline

Introduction

Introduction to Reinforcement Learning

- We now introduce the framework of *reinforcement learning* (RL), which encompasses a rich set of dynamic, stateful decision making problems.
- In the language of bandits, for each decision π^t, it is now a *multi-stage strategies*, rather than one-shot decision. To be specific, for each time (which is now termed as episode) $t = 1, 2, \ldots, T$, the learner acts for H steps.
- Another characteristic of RL is that the environments can now have multiple states and transitions, which will depend on the actions and state of previous state of the environment.
- While the action in RL will depend on the state of environment, the reward will depend on both the action chosen and the state of the environment.
- RL includes many of the previous bandit problems as special cases as will be seen later.

Finite-Horizon Episodic MDP Formulation

Markov Decision Process

In RL, the interactions between the agent and the environment for a single episode are often described by an (in)finite-horizon, Markov Decision Process (MDP): $M = (\mathcal{S}, \mathcal{A}, \{P_h^M\}_{h=1}^H, \{R_h^M\}_{h=1}^H, d_1)$ specified by:

- A state space $\mathcal{S}$, which may be finite (or infinite).
- An action space $\mathcal{A}$, which also may be discrete or infinite.
- A state-dependent transition function $P_h^M : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, where $\Delta(\mathcal{S})$ is the space of probability distributions over $\mathcal{S}$.
  - $P_h(s'|s, a)$ is the probability of transitioning into state s' upon taking action a in state s at step h.
- A time-dependent reward function $R_h^M : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$, the immediate reward associated with taking action a in state s at step h.
- The integer H which defines the horizon of the problem.
- An initial state distribution d_1, which specifies how the initial state s_1 is generated.

Markov Decision Process

For any fixed MDP M, an episode proceeds under the following protocol.

1. At the beginning of the episode, the learner selects a randomized, non-stationary (i.e. different across different steps) policy

$$\pi = (\pi_1, \ldots, \pi_H),$$

   where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$, Π_{rns} is the collection for all "randomized, non-stationary" policies.

2. The episode then evolves through the following process, beginning from $s_1 \sim d_1$. For $h = 1, \ldots, H$:
   - $a_h \sim \pi_h(s_h)$,
   - $r_h \sim R_h^M(s_h, a_h)$ and $s_{h+1} \sim P_h^M(s_h, a_h)$.

- For convenience, s_{H+1} is set to be a deterministic terminal state.
- The Markov property refers to the fact that

$$P_h^M(s_{h+1} = s' | s_h, a_h) = P_h^M(s_{h+1} = s' | s_h, a_h, s_{h-1}, a_{h-1}, \ldots, s_1, a_1).$$

Values and Goals of RL

- The value for a policy π under M is given by

$$f^M(\pi) := \mathbb{E}^{M,\pi}\left[\sum_{h=1}^{H} r_h\right],$$

where $\mathbb{E}^{M,\pi}$ denotes expectation under the process above, with respect to the randomness of state transitions and the stochasticity of π, and possibly the reward function (which however will be assumed to be deterministic later).

- The optimal policy for model M is defined as

$$\pi_M = \arg\max_{\pi \in \Pi_{\text{rns}}} f^M(\pi). \tag{1}$$

Values and Goals of RL

- Maximization in (1) is a daunting task, since each policy π is a complex multi-stage object consisting of H steps.
- To facilitate analysis, we break this complex task into smaller sub-tasks.
- Specifically, for a given model M and policy π, we define the state-action value function and state value function via

$$Q_h^{M,\pi}(s,a) = \mathbb{E}^{M,\pi}\left[\sum_{h'=h+1}^{H} r_{h'} \mid s_h = s, a_h = a\right], \tag{2}$$

$$V_h^{M,\pi}(s) = \mathbb{E}^{M,\pi}\left[\sum_{h'=h}^{H} r_{h'} \mid s_h = s\right]. \tag{3}$$

- Hence, the definition in (1) reads

$$f^M(\pi) = \mathbb{E}_{s\sim d_1, a\sim\pi_1(s)}[Q_1^{M,\pi}(s,a)] = \mathbb{E}_{s\sim d_1}[V_1^{M,\pi}(s)] \tag{4}$$

Online RL and the Regret

- We will focus on online RL problem that interacts with an unknown MDP M^* for T episodes. For each episode $t = 1, \ldots, T$, the learner selects a policy $\pi^t \in \Pi_{\text{rns}}$ and could observe the following trajectory

$$\tau^t = (s_1^t, a_1^t, r_1^t), \ldots, (s_H^t, a_H^t, r_H^t).$$

- The goal is to minimize the total regret

$$\sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p_t}[f^{M^*}(\pi^{M^*}) - f^{M^*}(\pi^t)] \tag{5}$$

  against the optimal policy π^{M^*} for M^*.
- As can be seen, Online RL is a strict generalization of (structured) bandits and contextual bandits (with i.i.d. contexts)
  - if $\mathcal{S} = \{s_0\}$ and $H = 1$, each episode amounts to choosing an action $a \in \mathcal{A}$ and observing a reward r^t with mean $f^M(a^t)$, which is precisely a bandit problem
  - taking $\mathcal{S} = X$ and $H = 1$ puts us in the setting of contextual bandits, with d_1 being the distribution of contexts

Planning via Dynamic Programming

Panning via Dynamic Programming

- To bound the regret in (5), we need to understand the structure of solutions to (1) in the case where M^* is known to the decision-maker.
- It will be shown that the problem of solving (1) for known M (known as *planning*) can be solved efficiently via the principle of *dynamic programming*, which solves a complex multi-stage decision (policy) by breaking down it into a sequence of small decisions.
- A fundamental result in dynamic programming is the existence of an optimal policy $\pi_M = (\pi_{M,1}, \dots, \pi_{M,H})$ that maximizes $V_1^{M,\pi}(s)$ over Π_{rns} for all states $s \in \mathcal{S}$ simultaneously (rather than just on average, as in (1))
- The intuition for such a results is that if $\pi_{M,h}(s)$ is defined for all $s \in \mathcal{S}$ and $h = 2, \dots, H$, then defining the optimal $\pi_{M,1(s)}$ at any state s to greedily choose an action that maximizes the sum of the expected immediate reward and the remaining expected reward under the optimal policy.

Optimal Value Functions

- To state the result formally, we introduce the optimal value functions:

$$\begin{aligned} Q_h^{M,*}(s,a) &= \max_{\pi \in \Pi_{\text{rns}}} \mathbb{E}^{M,\pi}\left[\sum_{h'=h}^{H} r_{h'} \mid s_h = s, a_h = a\right], \\ V_h^{M,*}(s) &= \max_a Q_h^{M,*}(s,a) \end{aligned} \tag{6}$$

  for all $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $h \in [H]$.
- We adopt the convention that $V_{H+1}^{M,*}(s) = Q_{H+1}^{M,*}(s,a) = 0$.
- It can be shown that there exists π_M such that for all s, a, h:

$$Q_h^{M,*}(s,a) = Q_h^{M,\pi_M}(s,a), \quad \text{and} \quad V_h^{M,*}(s) = V_h^{M,\pi_M}(s). \tag{7}$$

- A proof could be found in Theorem 1.7 in Reinforcement Learning: Theory and Algorithms, which is complicated but followable.

Proposition 1 (Bellman Optimality)

The optimal value functions in (6) for MDP M can be computed via $V_{H+1}^{M,\pi_M}(s) := 0$, and for each $s \in \mathcal{S}$,

$$V_h^{M,\pi_M}(s) = \max_{a \in \mathcal{A}} \mathbb{E}\left[r_h + V_{h+1}^{M,\pi_M}(s_{h+1}) \mid s_h = s, a_h = a\right]. \tag{8}$$

The optimal policy is given by:

$$\pi_{M,h}(s) \in \arg\max_{a \in \mathcal{A}} \mathbb{E}\left[r_h + V_{h+1}^{M,\pi_M}(s_{h+1}) \mid s_h = s, a_h = a\right]. \tag{9}$$

Equivalently, for all $s \in \mathcal{S}$, $a \in \mathcal{A}$,

$$Q_h^{M,\pi_M}(s,a) = \mathbb{E}\left[r_h + \max_{a' \in \mathcal{A}} Q_{h+1}^{M,\pi_M}(s_{h+1}, a') \mid s_h = s, a_h = a\right]. \tag{10}$$

and the optimal policy is given by

$$\pi_{M,h}(s) \in \arg\max_{a \in \mathcal{A}} Q^h_{M,\pi^M}(s,a). \tag{11}$$

Proof

- We only provide proof for (8), and all the others will follow by definition and noticing that $V_h^{M,\pi_M}(s) = \max_a Q_h^{M,\pi_M}(s,a)$.
- Proof of (8):

$$
\begin{aligned}
V_h^{M,\pi_M}(s) &= \max_{\pi\in\Pi} \mathbb{E}^M \left[r_h + \sum_{h'=h+1}^{H} r_{h'}(s_{h'}, a_{h'}) | s_h = s \right] \\
&= \max_{\pi\in\Pi} \mathbb{E}^M \left[r_h + \mathbb{E}\left[\sum_{h'=h+1}^{H} r_{h'}(s_{h'}, a_{h'}) \mid \pi, s_h = s, s_{h+1} \right] | s_h = s, \right] \\
&\le \max_{\pi\in\Pi} \mathbb{E}^M \left[r_h + \max_{\pi'\in\Pi} \mathbb{E}\left[\sum_{h'=h+1}^{H} r_{h'}(s_{h'}, a_{h'}) \mid \pi', s_h = s, s_{h+1} \right] | s_h = s, \right] \\
&= \max_{\pi\in\Pi} \mathbb{E}^M \left[r_h + V_{h+1}^{M,\pi_M}(s_{h+1}) | s_h = s \right] \\
&= \max_{a\in\mathcal{A}} \mathbb{E}^M \left[r_h + V_{h+1}^{M,\pi_M}(s_{h+1}) | s_h = s, a_h = a \right]
\end{aligned}
$$

The result then follow by definition that $V_h^{M,\pi_M}(s)$ is also the optimal value function.

Bellman Operator

- The update in (11) is referred to as *value iteration* (VI).
- We now define *Bellman Operators* which will be useful. For an MDP M, define the $\mathcal{T}_1^M, \ldots, \mathcal{T}_H^M$ via

$$\mathcal{T}_h^M Q(s, a) = \mathbb{E}^M \left[r_h(s_h, a_h) + \max_{a'} Q(s_{h+1}, a') \mid s_h = s, a_h = a \right]$$

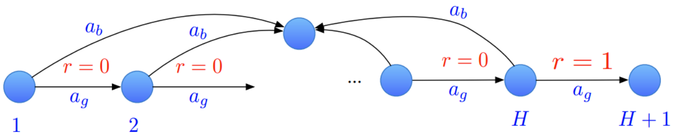
- In the language of Bellman operators, (11) can be written as

$$Q_h^{M,\pi_M} = \mathcal{T}_h^M Q_{h+1}^{M,\pi_M}$$

Failure of Uniform Exploration

Failure of Uniform Exploration

- Since we are interested in learning to make decisions in the face of an unknown environment, we will need exploration to minimize regret.
- While the ϵ-Greedy works for bandits and contextual bandits, albeit with a suboptimal rate ($T^{2/3}$ as opposed to $\sqrt{T}$), it can be disastrous in RL as it leads to exponential (in the horizon H) regret.
- "combination lock" example:



"Combination Lock" Example

- There are $H + 2$ states, and two actions a_g and a_b, and a starting state 1.
- The "good" action a_g deterministically leads to the next state in the chain, while the "bad" action deterministically leads to a terminal state.
- The only place where a non-zero reward can be received is the last state H, if the good action is chosen.
- So the only way to receive non-zero reward is to select a_g for all the H steps within the episode.
- Since the length of the episode is also H, selecting actions uniformly brings no information about the optimal sequence of actions, unless by chance all of the actions sampled happen to be good;
- The probability that this occurs is exponentially small in H.
- This means that T needs to be at least $O(2^H)$ to achieve nontrivial regret, and highlights the need for more strategic exploration.

UCB Methods and the Anaysis Tools

UCB Methods and the Analysis Tools

- While ϵ-Greedy method fails, it can be shown that UCB method yields a regret bound that is polynomial in the parameters $|\mathcal{S}|, |\mathcal{A}|$, and H.
- We now introduce two very important lemmas:
  1. The Performance Difference Lemma: that expresses the difference in values for two policies in terms of differences in single-step decisions made by the two policies.
  2. The Bellman residual decomposition: that relates the performance of the same policy under two different MDPs.

Lemma 1 (Performance Difference Lemma)

For any $s \in \mathcal{S}$, *and* $\pi, \pi' \in \Pi_{rns}$,

$$V_1^{M,\pi'}(s) - V_1^{M,\pi}(s) = \sum_{h=1}^{H} \mathbb{E}^{M,\pi}\left[Q_h^{M,\pi'}(s_h, \pi'(s_h)) - Q_h^{M,\pi'}(s_h, a_h) \mid s_1 = s\right]$$

- The proof proceeds by successively changing one policy into another and keep track of the ensuing differences in expected rewards.
- *Proof*: Fix a pair of policies π, π' and define $\pi^h = (\pi_1, \ldots, \pi_{h-1}, \pi'_h, \ldots, \pi'_H)$, with $\pi^1 = \pi'$ and $\pi^H = \pi$. By telescoping, we can write

$$V_1^{M,\pi'}(s) - V_1^{M,\pi}(s) = \sum_{h=1}^{H}\left[V_1^{M,\pi^h}(s) - V_1^{M,\pi^{h+1}}(s)\right] \tag{12}$$

Observe that for each h, we have

$$V_1^{M,\pi^h}(s) - V_1^{M,\pi^{h+1}}(s) = \mathbb{E}^{M,\pi^h}\left[\sum_{h=1}^{H} r_h \mid s_1 = s\right] - \mathbb{E}^{M,\pi^{h+1}}\left[\sum_{h=1}^{H} r_h \mid s_1 = s\right] \tag{13}$$

Here, one process evolves according to (M, π^h) and the one evolves according to (M, π^{h+1}). The processes only differ in the action taken once the state s_h is reached. In the former, the action $\pi'(s_h)$ is taken, whereas in the latter it is $\pi(s_h)$. Hence, equation (13) is equal to

$$\mathbb{E}^{M,\pi}\left[Q_h^{M,\pi'}(s_h, \pi'(s_h)) - Q_h^{M,\pi'}(s_h, \pi(s_h)) \mid s_1 = s\right] \tag{14}$$

which can be written as

$$\mathbb{E}^{M,\pi}\left[Q_h^{M,\pi'}(s_h, \pi'(s_h)) - Q_h^{M,\pi'}(s_h, a_h) \mid s_1 = s\right]. \tag{15}$$

Lemma 2 (Bellman residual decomposition)

For any pair of MDPs $M = (P^M, R^M)$ *and* $\tilde{M} = (\tilde{P}^M, \tilde{R}^M)$, *for any* $s \in \mathcal{S}$, *and policies* $\pi \in \Pi_{RNS}$,

$$V_1^{M,\pi}(s) - V_1^{\hat{M},\pi}(s) = \sum_{h=1}^{H} \mathbb{E}^{\hat{M},\pi}\left[Q_h^{M,\pi}(s_h, a_h) - r_h - V_{h+1}^{M,\pi}(s_{h+1}) \mid s_1 = s\right] \tag{16}$$

Hence, for $M, \hat{M}$ *with the same initial state distribution,*

$$f^M(\pi) - f^{\hat{M}}(\pi) = \sum_{h=1}^{H} \mathbb{E}^{M,\pi}\left[Q_h^{M,\pi}(s_h, a_h) - r_h - V_{h+1}^{M,\pi}(s_{h+1})\right]. \tag{17}$$

In addition, for any MDP M *and function* $Q = (Q_1, \ldots, Q_H, Q_{H+1})$ *with* $Q_{H+1} = 0$, *letting* $\pi_{Q,h}(s) = \arg\max_{a \in A} Q_h(s, a)$, *we have*

$$\max_{a \in A} Q_1(s, a) - V_1^{M,\pi_Q}(s) = \sum_{h=1}^{H} \mathbb{E}^{M,\pi_Q}\left[Q_h(s_h, a_h) - \mathcal{T}_h^M Q_{h+1}(s_h, a_h) \mid s_1 = s\right]. \tag{18}$$

and, hence,

$$\mathbb{E}_{s_1 \sim d_1}\left[\max_{a \in A} Q_1(s_1, a) - f^M(\pi_Q)\right] = \sum_{h=1}^{H} \mathbb{E}^{M,\pi_Q}\left[Q_h(s_h, a_h) - \mathcal{T}_h^M Q_{h+1}(s_h, a_h)\right]. \tag{19}$$

Proof I

We will prove (17), and omit the proof for (16), which is similar but more verbose. We have:

$$
\begin{aligned}
& \sum_{h=1}^{H} \mathbb{E}^{\hat{M},\pi}\left[Q_h^{\hat{M},\pi}(s_h, a_h) - r_h - V_{h+1}^{M,\pi}(s_{h+1})\right] \\
= \quad & \sum_{h=1}^{H} \mathbb{E}^{\hat{M},\pi}\left[Q_h^{M,\pi}(s_h, a_h) - V_{h+1}^{M,\pi}(s_{h+1})\right] - \mathbb{E}^{\hat{M},\pi}\left[\sum_{h=1}^{H} r_h\right] \\
= \quad & \sum_{h=1}^{H} \mathbb{E}^{\hat{M},\pi}\left[Q_h^{M,\pi}(s_h, a_h) - V_{h+1}^{M,\pi}(s_{h+1})\right] - f^{\hat{M}}(\pi).
\end{aligned}
$$

Proof II

On the other hand, since $V_h^{M,\pi}(s) = \mathbb{E}_{a\sim\pi_h(s)}[Q_h^{M,\pi}(s,a)]$, a telescoping argument yields

$$\begin{aligned} & \sum_{h=1}^{H} \mathbb{E}^{\hat{M},\pi}[Q_h^{M,\pi}(s_h,a_h) - V_{h+1}^{M,\pi}(s_{h+1})] \\ = & \sum_{h=1}^{H} \mathbb{E}^{\hat{M},\pi}[V_h^{M,\pi}(s_h) - V_{h+1}^{M,\pi}(s_{h+1})] \\ = & \ \mathbb{E}^{\hat{M},\pi}[V_1^{M,\pi}(s_1)] - \mathbb{E}^{\hat{M},\pi}[V_{H+1}^{M,\pi}(s_{H+1})] = f^M(\pi), \end{aligned}$$

where we have used that $V_{M,\pi}^{H+1} = 0$, and that both MDPs have the same initial state distribution.

Proof III

We prove (19) (omitting the proof of (18)) using a similar argument. We have

$$
\begin{aligned}
& \sum_{h=1}^{H} \mathbb{E}^{M,\pi_Q}[Q_h(s_h, a_h) - r_h - \max_{a' \in \mathcal{A}} Q_{h+1}(s_{h+1}, a')] \\
= & \sum_{h=1}^{H} \mathbb{E}^{M,\pi_Q}[Q_h(s_h, a_h) - \max_{a \in \mathcal{A}} Q_{h+1}(s_{h+1}, a) - \mathbb{E}^{M,\pi_Q}\left[\sum_{h=1}^{H} r_h\right], \\
= & \sum_{h=1}^{H} \mathbb{E}^{M,\pi_Q}[Q_h(s_h, a_h) - \max_{a \in \mathcal{A}} Q_{h+1}(s_{h+1}, a) - f^M(\pi_Q)].
\end{aligned}
$$

Since $a_{h+1} = \pi_{Q,h}(s_{h+1}) = arg \max_{a \in \mathcal{A}} Q_{h+1}(s_{h+1}, a)$, we have $\mathbb{E}^{M,\pi_Q}[Q_h(s_h, a_h) - \max_{a \in \mathcal{A}} Q_{h+1}(s_{h+1}, a)] = \mathbb{E}^{M,\pi_Q}[Q_h(s_h, a_h) - Q_{h+1}(s_{h+1}, a_{h+1})]$, the result follows by telescoping.

Optimism

Error Decomposition for Optimistic Policies

- We now turn back to the development of UCB algorithm for RL.
- Before constructing a sequence of optimistic *value functions* $\bar{Q}_1, \dots, \bar{Q}_H$, which are guaranteed to over-estimate the optimal value function Q^*_M, we first introduce the following lemma

Lemma 3 (Error Decomposition for Optimistic Policies)

Let $\{\bar{Q}_1, \dots, \bar{Q}_H\}_{h=1}^H$ *be a sequence of functions* $\bar{Q}_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ *with the property that for all* (s, a),

$$Q_h^{M,*}(s, a) \leq \bar{Q}_h(s, a) \tag{20}$$

and set $\bar{Q}_{H+1} = 0$*. Let* $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_H)$ *be such that* $\hat{\pi}_h(s) = \arg\max_a \bar{Q}_h(s, a)$*. Then for all* $s \in S$,

$$V_1^{M,*}(s) - V_1^{M,\hat{\pi}}(s) \leq \sum_{h=1}^{H} \mathbb{E}^{M,\hat{\pi}} \left[(\bar{Q}_h - \mathcal{T}_h^M \bar{Q}_{h+1})(s_h, \hat{\pi}(s_h)) \mid s_1 = s \right]. \tag{21}$$

Proof

- The lemma tells us that closeness of $\bar{Q}_h$ to the Bellman backup $\mathcal{T}_h^M \bar{Q}_{h+1}$ implies closeness of $\hat{\pi}$ to π_M in terms of the value.
- *Proof*: Let $\bar{V}_h(s) = \max_a \bar{Q}_h(s, a)$. Just as in the proof of Lemma 7, the assumption that $\bar{Q}_h$ is "optimistic" implies that

$$Q_h^{M,*}(s_h, \pi_M(s_h)) \le \bar{Q}_h(s_h, \pi_M(s_h)) \le \bar{Q}_h(s_h, \hat{\pi}_M(s_h)),$$

and hence, $V_1^{M,*}(s) \le \bar{V}_1(s)$. Then, (18) applied to $Q = \bar{Q}$ and $\pi_Q = \hat{\pi}$ states that

$$\bar{V}_1(s) - V_1^{M,\hat{\pi}}(s) = \sum_{h=1}^{H} \mathbb{E}_{M,\hat{\pi}}[(\bar{Q}_h(s_h, a_h) - \mathcal{T}_h^M \bar{Q}_{h+1}(s_h, a_h)) \mid s_1 = s].$$

The UCB-VI Algorithm for Tabular MDPs

The UCB-VI Algorithm for Tabular MDPs

- We now instantiate the principle of optimism to give regret bounds for online RL in *tabular MDPs*, where the state and action spaces are small (or finite).
- For simplicity, we assume that the reward function is known to the learner, so that only the transition probabilities are unknown.
- We will show that the regret bounds we present will depend polynomially on $|\mathcal{S}|$ and $|\mathcal{A}|$, as well as the horizon H.
- Define, with a slight abuse of notation,

$$n_t(s,a) = \sum_{i=1}^{t-1} \mathbf{1}((s_i,a_i)=(s,a)), \ \ n_t(s,a,s') = \sum_{i=1}^{t-1} \mathbf{1}((s_i,a_i,s_{i+1})=(s,a,s')),$$

We can estimate the transition probabilities via

$$\hat{P}_t(s'|s,a) = \frac{n_t(s,a,s')}{n_t(s,a)}. \tag{22}$$

The UCB-VI Algorithm

The following algorithm, UCB-VI (*Upper Confidence Bound Value Iteration*), combines the notion of optimism with dynamic programming.

The UCB-VI algorithm. The following algorithm, UCB-VI ("Upper Confidence Bound Value Iteration") [16], combines the notion of optimism with dynamic programming.

UCB-VI
for $t = 1, \ldots, T$ **do**
  Let $\overline{V}^t_{H+1} \equiv 1$.
  for $h = H, \ldots, 1$ **do**
    Update $n^t_h(s,a)$, $n^t_h(s,a,s')$, and $b^t_{h,\delta}(s,a)$, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.
    `//` $b^t_{h,\delta}(s,a)$ `is a bonus computed in (5.27).`
    Compute:

$$\overline{Q}^t_h(s,a) = \left\{ r_h(s,a) + \mathbb{E}_{s' \sim \widehat{P}^t_h(\cdot|s,a)} \overline{V}^t_{h+1}(s') + b^t_{h,\delta}(s,a) \right\} \wedge 1. \tag{5.26}$$

    Set $\overline{V}^t_h(s) = \max_{a \in \mathcal{A}} \overline{Q}^t_h(s,a)$ and $\widehat{\pi}^t_h(s) = \arg\max_{a \in \mathcal{A}} \overline{Q}^t_h(s,a)$.
  Collect trajectory $(s^t_1, a^t_1, r^t_1), \ldots, (s^t_H, a^t_H, r^t_H)$ according to $\widehat{\pi}^t$.

Regret Bounds

- The bonus functions play precisely the same role as the width of the confidence interval in (2.19): these bonuses ensure that (20) holds with high probability, as we as ensuring $\bar{Q}_h$ to be "self-consistent" as required by (21).
- With an appropriate choice of bonus, the above algorithm achieves a polynomial regret bound.

Theorem 4

For any $\delta > 0$, *UCB-VI with*

$$b_{h,t}(s,a) = 2\sqrt{\frac{\log(2SAHT/\delta)}{n_t(s,a)}} \tag{23}$$

guarantees that with probability at least $1-\delta$,

$$Reg \lesssim HS\sqrt{AT} \cdot \sqrt{\log(SAHT/\delta)}$$

Analysis for a Single Episode

- To bound the regret for UCB-VI, we first focus on a single episode by fixing t and prove several useful lemmas.
- Given the estimated transitions $\hat{P}_h(.|s,a), \{\mathcal{S}, \mathcal{A}, \{\hat{P}_h^H\}, \{R_h^M\}, d_1\}$, the associated Bellman operator is

$$\mathcal{T}_h^{\hat{M}} Q(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim \hat{P}_h(.|s,a)}[\max_a Q(s',a)]$$

- Consider the sequence of functions $\bar{Q}_h : \mathcal{S} \times \mathcal{A} \to [0,1]$, where $\bar{V}_h : \mathcal{S} \to [0,1]$, with $\bar{Q}_{H+1} = 0$ and

$$\bar{Q}_h(s,a) = [\mathcal{T}_h^{\hat{M}} \bar{Q}_{h+1}(s,a) + b_h(s,a)] \wedge 1, \text{ and } \bar{V}_h(s) = \max_a \bar{Q}_h(s,a). \tag{24}$$

  for bonus functions $b_{h,\delta} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ to be chosen later.
- The following lemma shows that as long as the bonuses $b_{h,\delta}$ are large enough to bound the error between the estimated transition probabilities and true transition probabilities, the functions $\overline{Q}_1, \ldots, \overline{Q}_H$ constructed above will be optimistic.

Lemma 5

Suppose we have estimates $\hat{P}_h(. \mid s, a)$ *and a function* $b_{h,\delta}(s, a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ *with the property that for all* $s \in \mathcal{S}$, $a \in \mathcal{A}$,

$$\sum_{s'} \hat{P}_h(s' \mid s, a) V_h^{M,*}(s') - \sum_{s'} P_h^M(s' \mid s, a) V_h^{M,*}(s') \leq b_{h,\delta}(s, a).$$

Then for all $h \in [H]$, *we have*

$$\bar{Q}_h \geq Q_h^{M,*}, \quad \text{and} \quad \bar{V}_h \geq V_h^{M,*},$$

for $\bar{Q}_h, \bar{V}_h$ *defined in (24).*

Proof. *The proof proceeds by backward induction on the statement*

$$\bar{V}_h \geq V_h^{M,*},$$

with $h = H + 1$ *down to* $h = 1$. *We start with the base case* $h = H + 1$, *which is trivial because* $V_{H+1}^* = 0$. *Now, assume* $V_{h+1}^* \leq \bar{V}_{h+1}$, *and let us* *prove the induction step. Fix* $s, a \in \mathcal{S} \times \mathcal{A}$. *If* $\bar{Q}_h(s,a) = 1$, *then trivially,* $\bar{Q}_h(s,a) \geq Q_h^{M,*}(s,a)$. *Otherwise,* $\bar{Q}_h(s,a) = \mathcal{T}_h^{\hat{M}} \bar{Q}_{h+1}(s,a) + b_{h,\delta}(s,a)$, *and thus*

$$\begin{aligned} & \bar{Q}_h(s,a) - Q_h^{M,*}(s,a) \\ = \; & b_{h,\delta}(s,a) + \mathbb{E}_{s' \sim \hat{P}_h(\cdot|s,a)}[\bar{V}_{h+1}(s')] - \mathbb{E}_{s' \sim P_h^M(\cdot|s,a)}[V_{h+1}^{M,*}(s')] \\ \geq \; & b_{h,\delta}(s,a) + \mathbb{E}_{s' \sim \hat{P}_h(\cdot|s,a)}[V_{h+1}^{M,*}(s')] - \mathbb{E}_{s' \sim P_h^M(\cdot|s,a)}[V_{h+1}^{M,*}(s')] \geq 0. \end{aligned}$$

which implies that

$$\bar{V}_h(s) = \max_a \bar{Q}_h(s,a) \geq \max_a Q_h^*(s,a) = V_h^{M,*}(s),$$

concluding the induction step.

- We now analyze the effect of using an estimated model $\hat{M}$ for the Bellman operator rather than the true unknown $\mathcal{T}_h^M$

Lemma 6

Suppose we have estimates $\hat{P}_h(. \mid s, a)$, and $b'_{h,\delta}(s, a)$ with the property that

$$\max_{v \in [0,1]^s} \left| \sum_{s'} \hat{P}_h(s' \mid s, a) v(s') - \sum_{s'} P_h^M(s' \mid s, a) v(s') \right| \leq b'_{h,\delta}(s, a),$$

then the Bellman residual satisfies

$$\bar{Q}_h - \mathcal{T}_h^M \bar{Q}_{h+1} \leq (b_{h,\delta} + b'_{h,\delta}) \wedge 1$$

for $\bar{Q}_h, \bar{V}_h$ defined in (24).

Proof

That $\bar{Q}_h - \mathcal{T}_h^M \bar{Q}_{h+1} \leq 1$ is immediate. To prove the main result, observe that

$$\bar{Q}_h - \mathcal{T}_h^M \bar{Q}_{h+1} = \left\{\mathcal{T}_h^M \bar{Q}_{h+1} + b_{h,\delta}\right\} \wedge 1 - \mathcal{T}_h^M \bar{Q}_{h+1} \leq (\mathcal{T}_h^{\hat{M}} - \mathcal{T}_h^M)\bar{Q}_{h+1} + b_{h,\delta}$$

For any $Q \in \mathcal{S} \times \mathcal{A} \to [0,1]$,

$$\begin{aligned}
(\mathcal{T}_h^{\hat{M}} - \mathcal{T}_h^M)Q(s,a) &= \mathbb{E}_{s' \sim \hat{P}_h(\cdot|s,a)}\left[\max_a Q(s',a)\right] - \mathbb{E}_{s' \sim P_h^M(\cdot|s,a)}\left[\max_a Q(s',a)\right] \\
&\leq \max_{V \in [0,1]^S} \left|\mathbb{E}_{s' \sim \hat{P}_h(\cdot|s,a)}[V(s')] - \mathbb{E}_{s' \sim P_h^M(\cdot|s,a)}[V(s')]\right|.
\end{aligned}$$

Since the maximum is achieved at a vertex of $[0,1]^S$, the statement follows.

Regret Analysis

- We now reintroduce the time index t and demonstrate that the estimated transition probabilities in UCB-VI satisfy conditions of Lemma 5 and Lemma 6, ensuring that the functions $\bar{Q}_1, \dots, \bar{Q}_H$ are optimistic.

Lemma 7

Let $\{\hat{P}_h^t\}$ be defined as in (22). Then with probability at least $1-\delta$, the functions

$$b_{h,t}(s,a) = 2\sqrt{\frac{\log(2SAHT/\delta)}{n_t^h(s,a)}}, \quad b'_{h,t}(s,a) = 8\sqrt{\frac{\log(2SAHT/\delta)}{n_t^h(s,a)}}$$

satisfy the assumptions of Lemma 5 and Lemma 6, respectively, for all $s \in \mathcal{S}$, $a \in \mathcal{A}$, $h \in [H]$, and $t \in [T]$ simultaneously.

- See Lemma 7.2 in Reinforcement Learning: Theory and Algorithms for a proof.

Proof of Theorem 1. I

Integrating the results from Lemma 7, the $\bar{Q}_1, \ldots, \bar{Q}_H$ are shown to be optimistic, meaning the conditions of Lemma 3 hold, and the instantaneous regret on round t (conditionally on $s_1 \sim d_1$) is at most:

$$
\begin{aligned}
& \sum_{h=1}^{H} \mathbb{E}^{M,\hat{\pi}_t}[(\bar{Q}_h - \mathcal{T}_h^M \bar{Q}_{h+1})(s_h, \pi_t(s_h)) \mid s_1 = s] \\
\leq \quad & \sum_{h=1}^{H} \mathbb{E}^{M,\hat{\pi}_t}[(b_{h,\delta}(s_h, \hat{\pi}_t(s_h)) + b'_{h,\delta}(s_h, \hat{\pi}_t(s_h))) \mid s_1 = s],
\end{aligned}
$$

where the second inequality invokes Lemma 6. Summing over $t = 1, \ldots, T$, and applying the Azuma-Hoeffding inequality, we have that with probability at least $1 - \delta$, the regret of UCB-VI is bounded by:

$$
\begin{aligned}
& \sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}^{M,\hat{\pi}^t} \left[b_{h,\delta}(s_h^t, \hat{\pi}_h^t(s_h^t)) + b'_{h,\delta}(s_h^t, \hat{\pi}_h^t(s_h^t)) \right] \wedge 1 \\
\leq \quad & \sum_{t=1}^{T} \sum_{h=1}^{H} \left(b_{h,\delta}(s_h^t, \hat{\pi}_h^t(s_h^t)) + b'_{h,\delta}(s_h^t, \hat{\pi}_h^t(s_h^t)) \right) \wedge 1 + \sqrt{HT \log(1/\delta)}.
\end{aligned}
$$

Proof of Theorem 1. II

Using the bonus definition in (23), the bonus term above is bounded by

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\sqrt{\frac{S\log(2SAHT/\delta)}{n_t^h(s_h^t,\hat{\pi}_h^t(s_h^t))}}\wedge 1 \le \sqrt{S\log(2SAHT/\delta)}\sum_{t=1}^{T}\sum_{h=1}^{H}\frac{1}{\sqrt{n_t^h(s_h^t,\hat{\pi}_h^t(s_h^t))}}\wedge 1.$$

The double summation can be handled in the same fashion as Lemma 8:

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\frac{1}{\sqrt{n_t^h(s_h^t,\hat{\pi}_h^t(s_h^t))}}\wedge 1 = \sum_{h=1}^{H}\sum_{(s,a)}\sum_{t=1}^{T}\frac{\mathbb{I}\{(s_h^t,\hat{\pi}_h^t(s_h^t))=(s,a)\}}{\sqrt{n_t^h(s,a)}}\wedge 1$$

$$\le \sum_{h=1}^{H}\sum_{(s,a)}\sqrt{n_h^T(s,a)} \le H\sqrt{SAT}.$$

Thank You !