Theoretical Machine Learning
Lecture 10

Instructor: Haipeng Luo

1 Partial Monitoring: Lower Bounds

Recall the classification theorem for partial monitoring we discussed last time:

Theorem 1. *The minimax regret of a partial monitoring problem G is*

$$\inf_{\text{learner}} \max_{z_{1:n}} \mathbb{E}\left[\text{Reg}_n\right] = \begin{cases} 0, & \textit{if G has only one Pareto-optimal action;} \\ \Theta(\sqrt{n}), & \textit{else if G is locally observable;} \\ \Theta(n^{\frac{2}{3}}), & \textit{else if G is globally observable;} \\ \Theta(n), & \textit{else.} \end{cases}$$

We have proven all the upper bounds, and now we continue to prove the lower bounds. Note that we have discussed that MAB, a locally observable problem, admits $\Omega(\sqrt{n})$ minimax regret. On the other hand, we also gave an example of a hopeless problem

$$\ell = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \text{and} \qquad \Phi = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

which is not globally observable and clearly is not learnable. It is in fact also not hard to come up with a globally (but not locally) observable problem where $\Omega(n^{\frac{2}{3}})$ regret is intuitively unavoidable. For example, consider the spam detection problem again with query cost $c > 1/2$:

$$\ell = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ c & c \end{pmatrix} \qquad \text{and} \qquad \Phi = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 2 \end{pmatrix},$$

Recall that the only way to obtain information is by playing the query action. Therefore, if the environment selects one of the outcomes with probability $0.5 + \epsilon$ each time (and the other one with probability $0.5 - \epsilon$), then the learner either needs to query $1/\epsilon^2$ times to figure out which outcome appears more often, in which case she suffers $(c - (0.5 - \epsilon))/\epsilon^2 = \Omega(1/\epsilon^2)$ regret, or she does not query enough and never figures out which one is better, in which case she suffers $\Omega(n\epsilon)$ regret. By setting $\epsilon = n^{-\frac{1}{3}}$, we thus know that in either case the regret is at least $\Omega(n^{\frac{2}{3}})$.

So we have examples in each category with a matching regret lower bound. However, the classification theorem says something more — indeeds, it says that for *every* problem in each category, the corresponding regret lower bound is unavoidable for every algorithm. We provide a rigorous proof below, using similar ideas from the proof for the MAB lower bound.

Proof for lower bounds of Theorem 1. The idea is still to first construct a stochastic environment where two actions a and b are equally good, identify the one that is selected less often by the algorithm, and then construct another stochastic environment where this action becomes slightly better but it is hard for the algorithm to realize the change. Again, it is sufficient to consider deterministic algorithms. Below we first describe the common part of the proof for all three categories.

Let a and b be a pair of neighboring Pareto-optimal actions (we will specify which pair later for each category). Consider an environment where the outcomes $z_1, \ldots, z_n$ are i.i.d. samples from a distribution $u \in \Delta(d)$ that lies in the interior of $C_a \cap C_b$ (so by definition a and b are both optimal actions in this environment). For any fixed algorithm, let $m_k = \mathbb{E}\left[\sum_{t=1}^n \mathbf{1}\{a_t = k\}\right]$ be the expected total number of times action k is selected in this environment. Without loss of generality, assume $m_b \le n/2$ (note that one of m_a and m_b must be at most $n/2$).

Next consider a different environment where the outcomes $z_1, \ldots, z_n$ are i.i.d. samples from a distribution $u' = u + \delta$ where δ satisfies $\sum_k \delta(k) = 0$ and $\langle \ell_a - \ell_b, \delta \rangle = \epsilon$ for some small enough $\epsilon > 0$ such that $u' \in C_b$ (so action b is optimal under this new environment). Note that this is always possible since the constraint $\sum_k \delta(k) = 0$ defines a space that is orthogonal to the all-one vector, and $\ell_a - \ell_b$ cannot be in the same direction as the all-one vector for otherwise one of them strictly dominates the other.

It remains to argue that the regret of the same algorithm under this environment has to be large. Note that every time the algorithm selects action a, it incurs regret $\langle \ell_a - \ell_b, u' \rangle = \langle \ell_a - \ell_b, \delta \rangle = \epsilon$; and every time it selects an action $k \notin N_{ab}$, it incurs some constant regret which can be assumed to be larger than $c + \epsilon$ for some constant c as long as ϵ is small enough. Therefore, the regret is

$$\mathbb{E}'\left[\mathrm{Reg}_n\right] \ge \mathbb{E}\left[\sum_{t=1}^n \ell(a_t, z_t) - \sum_{t=1}^n \ell(b, z_t)\right] \ge (n - m'_b)\epsilon + c\bar{m}'$$

where $\mathbb{E}'$ denotes the expectation in environment u', $m'_k = \mathbb{E}'\left[\sum_{t=1}^n \mathbf{1}\{a_t = k\}\right]$, and $\bar{m}' = \sum_{k \notin N_{ab}} m'_k$.[1] Now we relate m_b and m'_b in a similar way as in the MAB lower bound proof:

$$m'_b \le m_b + n\left\|\mathbb{P} - \mathbb{P}'\right\|_1 \le m_b + n\sqrt{2\mathrm{KL}(\mathbb{P}' \parallel \mathbb{P})}$$

where $\mathbb{P}$ and $\mathbb{P}'$ are the distributions of the observation sequence $\Phi(a_1, z_1), \ldots, \Phi(a_n, z_n)$ in environment u and u' respectively. Generalizing the divergence decomposition lemma (Lemma 2 of Lecture 8), one can verify that

$$\mathrm{KL}(\mathbb{P}' \parallel \mathbb{P}) = \sum_{k=1}^K m'(k)\mathrm{KL}(\mathbb{P}'_k \parallel \mathbb{P}_k)$$

where $\mathbb{P}_k$ and $\mathbb{P}'_k$ are the distributions of $\Phi(k, z)$ when z is drawn from u and u' respectively, which, with our notation of signal matrix S_k, can be conveniently written as $S_k u$ and $S_k u'$. We next discuss the three categories separately.

Locally observable problems. It is not hard to verify that $\mathrm{KL}(\mathbb{P}'_k \parallel \mathbb{P}_k) = \mathcal{O}\left(\epsilon^2\right)$ (details omitted) for all k. Therefore

$$m'_b \le m_b + \mathcal{O}\left(\epsilon n\sqrt{n}\right) \le \frac{n}{2} + \mathcal{O}\left(\epsilon n\sqrt{n}\right),$$

and thus

$$\mathbb{E}'\left[\mathrm{Reg}_n\right] \ge (n - m'_b)\epsilon \ge \left(\frac{n}{2} - \mathcal{O}\left(\epsilon n\sqrt{n}\right)\right)\epsilon,$$

which is $\Omega(\sqrt{n})$ by setting $\epsilon = \beta/\sqrt{n}$ for some small enough constant β.

Non globally observable problems. In this case we let a and b be a neighboring pair that is not globally observable, and also let δ be orthogonal to rowspace$(S_{[K]})$. Note that this is always possible: the condition $\langle \ell_a - \ell_b, \delta \rangle = \epsilon$ can be satisfied since by definition $\ell_a - \ell_b \notin$ rowspace$(S_{[K]})$, and the condition $\sum_k \delta(k) = 0$ holds automatically since the all-one vector is in rowspace$(S_{[K]})$. In this construction, for every k we have

$$\mathrm{KL}(\mathbb{P}'_k \parallel \mathbb{P}_k) = \mathrm{KL}(S_k u' \parallel S_k u) = \mathrm{KL}(S_k u + S_k\delta \parallel S_k u) = \mathrm{KL}(S_k u \parallel S_k u) = 0,$$

and thus $m'_b \le m_b \le n/2$. Therefore, $\mathbb{E}'\left[\mathrm{Reg}_n\right] \ge (n - m'_b)\epsilon \ge \frac{n\epsilon}{2} = \Omega(n)$.

[1] We have assumed that there are no degenerated/duplicate actions. However, it is not hard to verify that the same statement holds up to some constant in the general case.

Globally (but not locally) observable problems. In this case we let a and b be a neighboring pair that is not locally observable, and let δ be orthogonal to $\text{rowspace}(S_{N_{ab}})$, which is also always possible by the fact $\ell_a - \ell_b \notin \text{rowspace}(S_{N_{ab}})$ and that the all-one vector is in $S_{N_{ab}}$. Therefore, for $k \in N_{ab}$, again we have $\text{KL}(\mathbb{P}'_k \parallel \mathbb{P}_k) = \text{KL}(S_k u + S_k \delta \parallel S_k u) = \text{KL}(S_k u \parallel S_k u) = 0$, and for $k \notin N_{ab}$, $\text{KL}(\mathbb{P}'_k \parallel \mathbb{P}_k) = \mathcal{O}\left(\epsilon^2\right)$. Combing everything we have $m'_b \leq n/2 + \mathcal{O}\left(n\epsilon\sqrt{2\bar{m}'}\right)$ and

$$\mathbb{E}'\left[\text{Reg}_n\right] \geq (n - m'_b)\epsilon + c\bar{m}' \geq \frac{n\epsilon}{2} - \mathcal{O}\left(n\epsilon^2\sqrt{2\bar{m}'}\right) + c\bar{m}' \geq \frac{n\epsilon}{2} - \mathcal{O}\left(n^2\epsilon^4\right),$$

where the last step is by lower bounding the quadratic (in terms of $\sqrt{\bar{m}'}$) by its minimum. Finally setting $\epsilon = \beta n^{-\frac{1}{3}}$ for some small enough constant β proves $\mathbb{E}'\left[\text{Reg}_n\right] = \Omega(n^{\frac{2}{3}})$. □

2 Reinforcement Learning

Partial monitoring generalizes MAB in terms of the feedback model. In the rest of this lecture, we discuss a completely different direction of generalizing MAB, called *reinforcement learning*, where the key difference compared to MAB is that the learner is *stateful* and the reward/loss from the environment depends on not only the action taken by the learner, but also the current state the learner is at. More importantly, the learner's action influences the state transition.

There are many different formal setups for reinforcement learning. Here, we focus on a specific one: episodic Markov decision process, parameterized by a state space $\mathcal{X}$, an action space $\mathcal{A}$, the number of steps H in each episode, a sequence of reward functions $r_1, \ldots, r_H \in [0,1]^{\mathcal{X}\times\mathcal{A}}$, and finally a sequence of transition functions $\mathbb{P}_1, \ldots, \mathbb{P}_H \in \Delta(\mathcal{X})^{\mathcal{X}\times\mathcal{A}}$. The learning protocol proceeds as follows.

For each episode $t = 1, \ldots, n$:

- The learner is reset to some arbitrary state $x_1^t \in \mathcal{X}$.
- For each step $h = 1, \ldots .H$:
  - The learner takes an action $a_h^t \in \mathcal{A}$, receives reward $r_h(x_h^t, a_h^t)$, and observes the next state x_{h+1}^t drawn from $\mathbb{P}_h(\cdot|x_h^t, a_h^t)$.

For convention we have switched from losses to rewards, but it makes no real difference. We assume that $\mathcal{X}, \mathcal{A}$ and $r_1, \ldots, r_H$ are known to the learner, but not $\mathbb{P}_1, \ldots, \mathbb{P}_H$. Note that even though the reward functions are known (so the learner knows exactly which action admits the highest reward at each state), the problem is still highly non-trivial since the transition functions are unknown, and thus the leaner needs to figure out which sequence of actions leads to high expected reward by interacting with the environment and observing samples of the transitions. In a more general setup, reward functions are also unknown and each step the learner only receives a reward sample with mean $r_h(x_h^t, a_h^t)$.

MAB can be seen as a special case where there is only one state ($|\mathcal{X}| = 1$) and one step in each episode ($H = 1$), and the reward function is unknown. Incorporating the concept of states and state transitions allows one to model a much broader class of problems, such as training a robot to finish a certain task or an agent to play Atari games. The Markov property (that is, the fact that the next state only depends on the current state and action but not the previous ones) is not very restrictive, as long as one encodes enough information in the state representation.

2.1 Bellman equations

The goal of the learner is again to minimize regret against the best policy, but what is the best policy in this problem, assuming knowledge of the transition functions? It is not hard to convince yourself that this can be solved by dynamic programing. Specifically, let $V_h^\star(x)$ be the highest expected reward one can achieve starting from state x at step h of an episode and behaving optimally afterwards, and similarly $Q_h^\star(x, a)$ be the highest expected reward one can achieve starting from state x with action a taken at step h and then behaving optimally in the future. Then these quantities satisfy the following well-known *Bellman equations*:

$$\begin{aligned}
V_h^\star(x) &= \max_{a\in\mathcal{A}} Q_h^\star(x,a), \\
Q_h^\star(x,a) &= r_h(x,a) + \sum_{x'\in\mathcal{X}} \mathbb{P}_h(x'|x,a)V_{h+1}^\star(x') \triangleq (r_h + \mathbb{P}_h V_{h+1}^\star)(x,a), \qquad (1)\\
V_{H+1}^\star(x) &= 0, \ \ \forall x \in \mathcal{X}.
\end{aligned}$$

So all $V_h^\star(x)$ and $Q_h^\star(x,a)$ can be found via a straightforward backward calculation. The optimal policy should take action $\operatorname{argmin}_a Q_h^\star(x,a)$ when at state x and step h of an episode, and the optimal reward achievable for an episode starting at state x_1 is $V_1^\star(x_1)$. Therefore, we can define the (pseudo) regret of a learner as

$$\overline{\text{Reg}}_n = \mathbb{E}\left[\sum_{t=1}^n V_1^\star(x_1^t) - \sum_{t=1}^n\sum_{h=1}^H r_h(x_h^t, a_h^t)\right].$$

Suppose the learner deploys a policy π_t at time t, which is a collection of mappings $\pi_1^t, \ldots, \pi_H^t \in \mathcal{A}^{\mathcal{X}}$. Then the regret can also be conveniently written with similar notation:

$$\overline{\text{Reg}}_n = \mathbb{E}\left[\sum_{t=1}^n V_1^\star(x_1^t) - \sum_{t=1}^n V_1^{\pi_t}(x_1^t)\right]$$

where

$$\begin{aligned}
V_h^{\pi_t}(x) &= r_h(x, \pi_h^t(x)) + \sum_{x'\in\mathcal{X}} \mathbb{P}_h(x'|x,\pi_h^t(x))V_{h+1}^{\pi_t}(x') \triangleq (r_h + \mathbb{P}_h V_{h+1}^{\pi_t})(x,\pi_h^t(x)), \qquad (2)\\
V_{H+1}^{\pi_t}(x) &= 0, \ \ \forall x \in \mathcal{X}.
\end{aligned}$$

2.2 Optimistic Q-learning

So how should the learner ensure low regret without knowledge of the transition functions? There are roughly two types of algorithms: model-based and model-free. Model-based ones try to build a good estimate of the transition functions using observed samples, and then act according to an approximately optimal strategy computed based on these transition estimates. Note that a transition function has $|\mathcal{X}|^2|\mathcal{A}|$ parameters, and thus model-based algorithms need at least $|\mathcal{X}|^2|\mathcal{A}|$ space. On the other hand, model-free algorithms do not estimate or maintain the model. Instead, they either directly estimate the Q-value function $Q_h^\star$ or directly optimize over the policy space. In either case, usually only $\mathcal{O}(|\mathcal{X}||\mathcal{A}|)$ space is needed.

In this lecture we focus on a particular type of model-free algorithms, called *Q-learning*. As the name suggests, Q-learning works by maintaining an estimate Q_h for the optimal Q-value function $Q_h^\star$ (and similarly an estimate $V_h(x) = \max_a Q_h(x,a)$ for $V_h^\star$). With the estimate Q_h, at state x and step h of an episode, the algorithm simply takes action $\operatorname{argmax}_a Q_h(x,a)$. The high-level idea of maintaining this estimate is as follows. Note that each time after taking an action a at some state x and step h and observing the next state x', the quantity $r_h(x,a) + V_{h+1}(x')$ serves as a nearly-unbiased sample of $(r_h + \mathbb{P}_h V_{h+1}^\star)(x,a) = Q_h^\star(x,a)$ as long as V_{h+1} is close enough to $V_{h+1}^\star$. So a natural update to Q_h is

$$Q_h(x,a) \leftarrow (1-\alpha)Q_h(x,a) + \alpha(r_h(x,a) + V_{h+1}(x')) \qquad (3)$$

for some coefficient $\alpha \in [0,1]$. For example, suppose for the m-th time we do this update for the specific (x,a) pair, we set the coefficient to be $\alpha_m = 1/m$. Then by expanding the recursive update, one can see that $Q_h(x,a)$ is exactly the average of the quantities of the form $(r_h(x,a) + V_{h+1}(x'))$ observed in the past.

One issue in the above argument is that the update only makes sense if we start with a good estimate already, but at the beginning we almost always start with an inaccurate estimate. This suggests that the coefficient $\alpha_m = 1/m$ is not a right choice, and in particular it should be set slightly larger so that the algorithm gradually forgets about the initial phase and put more focus on recent updates. It turns out that the right choice is $\alpha_m = \frac{H+1}{H+m}$ and we defer the reason to the analysis in next section.

However, another (more important) issue is that the algorithm lacks sufficient exploration. Indeed, to see this, consider applying this algorithm to the special case of MAB (so $|\mathcal{X}| = 1$ and $H = 1$).

Algorithm 1: Optimistic Q-learning

```
Initialize $Q_h(x,a) = H$ and counter $m_h(s,a) = 0$ for all $x, a, h$.
Define coefficient $\alpha_m = \frac{H+1}{H+m}$ and bonus term $b_m = \frac{c}{\sqrt{m}}$.
For each episode $t = 1, \dots, n$:
    • Reset to state $x_1^t \in \mathcal{X}$.
    • For each step $h = 1, \dots .H$:
        – Take action $a_h^t = \mathrm{argmax}_{a \in \mathcal{A}} Q_h(x,a)$ and observe the next state $x_{h+1}^t$.
        – Let $m = m_h(s,a) \leftarrow m_h(s,a) + 1$.
        – Update
              $Q_h(x_h^t, a_h^t) \leftarrow (1-\alpha_m) Q_h(x_h^t, a_h^t) + \alpha_m \left( r_h(x_h^t, a_h^t) + V_{h+1}(x_{h+1}^t) + b_m \right).  \quad (4)$
        – Update $V_h(x_h) \leftarrow \min \{ H, \max_{a \in \mathcal{A}} Q_h(x_h, a) \}$.
```

It is not hard to see that the algorithm is simply picking whichever action with the largest empirical reward and never explores, which clearly leads to linear regret.

In light of the optimistic idea from the UCB algorithm, a natural fix to the exploration issue is to add some "bonus" term to $Q_h(x,a)$ so that with high probability it is a tight upper bound of $Q^\star(x,a)$. Similar to UCB, the bonus term should correspond to some deviation bound and be of order $1/\sqrt{m}$ where m is again the number of times (x,a) is encountered. In this way the bonus term encourages exploring a pair (x,a) if it is selected not often enough, and shrinks as the pair is encountered more and more often.

It turns out that one convenient way to incorporate this bonus term into the incremental update Equation (3) is by adding a term of order $1/\sqrt{m}$ to the quantity $(r_h(x,a) + V_{h+1}(x'))$. Combining these ideas, the final algorithm is included in Algorithm 1, which is taken from a very recent work by Jin et al. [2018].

2.3 Regret analysis

In this section, we proceed to prove the regret guarantee of Algorithm 1, summarized in the following theorem.

Theorem 2. *With an appropriate chosen constant c, the regret of Algorithm 1 is at most* $\mathcal{O}\left(\sqrt{n|\mathcal{X}||\mathcal{A}|H^5}\right)$.

To prove this theorem, we first show that just like the UCB index, Q_h is indeed an upper confidence bound of $Q_h^\star$, but at the same time it is not much larger than $Q_h^\star$. To make the notation clear, we let Q_h^t be the value of Q_h at the beginning of episode t, and similarly m_h^t be the value of the counter m_h at the beginning of episode t. Therefore, for some fixed t, h, x and a, if $m = m_h^t(x,a)$ is the number of times (x,a) pair has been executed at step h before episode t, and $t_1 < \cdots < t_m < t$ are the episode indices where these executions happened, then by expanding the recursive update of Equation (4) one can verify that

$$\begin{aligned} Q_h^t(x,a) &= Q_h^{t_m+1}(x,a) = (1-\alpha_m) Q_h^{t_m}(x,a) + \alpha_m \left(r_h(x,a) + V_{h+1}^{t_m}(x_{h+1}^{t_m}) + b_m \right) \\ &= (1-\alpha_m)(1-\alpha_{m-1}) Q_h^{t_{m-1}}(x,a) + (1-\alpha_m)\alpha_{m-1} \left(r_h(x,a) + V_{h+1}^{t_{m-1}}(x_{h+1}^{t_{m-1}}) + b_{m-1} \right) \\ &\quad + \alpha_m \left(r_h(x,a) + V_{h+1}^{t_m}(x_{h+1}^{t_m}) + b_m \right) \\ &= \cdots \\ &= \underbrace{\Pi_{j=1}^m (1-\alpha_j)}_{\triangleq \alpha_m^0} H + \sum_{i=1}^m \underbrace{\alpha_i \Pi_{j=i+1}^m (1-\alpha_j)}_{\triangleq \alpha_m^i} \left(r_h(x,a) + V_{h+1}^{t_i}(x_{h+1}^{t_i}) + b_i \right). \end{aligned} \tag{5}$$

Note that the coefficients $\alpha_m^0, \dots, \alpha_m^m$ defined above naturally satisfy $\sum_{i=0}^m \alpha_m^0 = 1$, so $Q_h^t(x,a)$ is a convex combination of the initial value H and m quantities of the form $r_h(x,a) + V_{h+1}^{t_i}(x_{h+1}^{t_i}) + b_i$.

As mentioned earlier, if we were to set $\alpha_j = 1/j$, then $\alpha_m^i = 1/m$ for $i \neq 0$ and the convex combination is a simple average, which is not desirable. However, with the particular choice of $\alpha_j = \frac{H+1}{H+j}$, one can verify that α_m^i is negligible for $i \lessapprox (1 - \frac{1}{H})m$, so the convex combination puts most weights on the recent $\frac{1}{H}$ fraction of updates, which turns out to be very important. For either choice, note that α_m^0 is 1 if $m = 0$ and 0 otherwise, meaning that the initial value H only plays a role when we have not even encountered (x, a) once yet.

Based on the discussion above, it is also intuitive and can be verified that (details omitted)

$$\beta_m \triangleq \sum_{i=1}^m \alpha_m^i b_i = c \sum_{i=1}^m \frac{\alpha_m^i}{\sqrt{i}} \in \left[\frac{c}{\sqrt{m}}, \frac{2c}{\sqrt{m}}\right], \quad \forall m \geq 1$$

since α_m^i is negligible unless i is large enough so that the denominator is of order $\Theta(\sqrt{m})$. Therefore, compared to the update of Equation (3) (which clearly can be seen as setting $b_m = 0$), the extra term we add to Q_h^t in the optimistic update Equation (4) is exactly β_m, which is of order $\Theta(1/\sqrt{m})$, similar to the UCB algorithm. However, also note that the effect of this extra term is in some sense recursive, since the term $V_{h+1}^{t_i}$ includes additional extra bonuses from the future steps as well.

We are now ready to compare Q_h^t and $Q_h^\star$. First, by the fact $\sum_{i=0}^m \alpha_m^0 = 1$ and Bellman equation (1) we rewrite $Q_h^\star(x, a)$ as

$$Q_h^\star(x, a) = \alpha_m^0 Q_h^\star(x, a) + \sum_{i=1}^m \alpha_m^i Q_h^\star(x, a) = \alpha_m^0 Q_h^\star(x, a) + \sum_{i=1}^m \alpha_m^i \left(r_h + \mathbb{P}_h V_{h+1}^\star\right)(x, a).$$

Comparing this with Equation (5) we have

$$\begin{aligned} Q_h^t(x, a) - Q_h^\star(x, a) &= \alpha_m^0 (H - Q_h^\star(x, a)) + \sum_{i=1}^m \alpha_m^i \left(V_{h+1}^{t_i}(x_{h+1}^{t_i}) - \mathbb{P}_h V_{h+1}^\star(x, a)\right) + \beta_m \\ &= \alpha_m^0 (H - Q_h^\star(x, a)) + \sum_{i=1}^m \alpha_m^i \underbrace{\left(V_{h+1}^{t_i}(x_{h+1}^{t_i}) - V_{h+1}^\star(x_{h+1}^{t_i})\right)}_{\triangleq \phi_{h+1}^{t_i}} \\ &\quad + \sum_{i=1}^m \alpha_m^i \left(V_{h+1}^\star(x_{h+1}^{t_i}) - \mathbb{P}_h V_{h+1}^\star(x, a)\right) + \beta_m, \end{aligned} \tag{6}$$

where the summands $V_{h+1}^\star(x_{h+1}^{t_i}) - \mathbb{P}_h V_{h+1}^\star(x, a)$ of the last summation constitutes a martingale difference sequence (since $x_{h+1}^{t_i} \sim \mathbb{P}_h(\cdot|x, a)$) and thus by some concentration argument and setting c to be of order roughly $\sqrt{H^3}$ we have with high probability (details omitted)

$$\sum_{i=1}^m \alpha_m^i \left(V_{h+1}^\star(x_{h+1}^{t_i}) - \mathbb{P}_h V_{h+1}^\star(x, a)\right) \in \left[-\frac{c}{\sqrt{m}}, \frac{c}{\sqrt{m}}\right]. \tag{7}$$

Combining the fact $\beta_m \geq \frac{c}{\sqrt{m}}$ and $H \geq Q_h^\star(x, a)$ we have

$$Q_h^t(x, a) - Q_h^\star(x, a) \geq \sum_{i=1}^m \alpha_m^i \phi_{h+1}^{t_i},$$

which via an induction on h can be shown to be nonnegative. Indeed, the base case is trivial, and assuming $Q_{h+1}^t(x, a) \geq Q_{h+1}^\star(x, a)$ for all t, x and a, we have $\phi_{h+1}^{t_i} = \min\left\{H, \max_a Q_{h+1}^{t_i}(x_{h+1}^{t_i}, a)\right\} - \max_a Q_{h+1}^\star(x_{h+1}^{t_i}, a) \geq 0$, and thus $Q_h^t(x, a) \geq Q_h^\star(x, a)$. This shows that Q_h^t and V_h^t are indeed upper bounds of $Q_h^\star$ and $V_h^\star$ respectively with high probability.

On the other hand, combining Equation (5) and Equation (7) we also have

$$Q_h^t(x, a) - Q_h^\star(x, a) \leq \alpha_m^0 H + \sum_{i=1}^m \alpha_m^i \phi_{h+1}^{t_i} + \frac{3c\mathbf{1}\left\{m > 0\right\}}{\sqrt{m}}. \tag{8}$$

We are now ready to prove Theorem 2. First define $\delta_h^t = V_h^t(x_h^t) - V_h^{\pi_t}(x_h^t)$ so that the regret is bounded by

$$\sum_{t=1}^n \left(V_1^\star(x_1^t) - V_1^{\pi_t}(x_1^t)\right) \leq \sum_{t=1}^n \left(V_1^t(x_1^t) - V_1^{\pi_t}(x_1^t)\right) = \sum_{t=1}^n \delta_1^t.$$

Our goal is to relate $\sum_{t=1}^n \delta_h^t$ to $\sum_{t=1}^n \delta_{h+1}^t$. To show this, note that

$$\begin{aligned}\delta_h^t &\le \max_{a\in\mathcal{A}} Q_h^t(x_h^t,a) - V_h^{\pi_t}(x_h^t)\\ &= Q_h^t(x_h^t,a_h^t) - V_h^{\pi_t}(x_h^t) && (a_h^t \in \mathrm{argmax}_{a\in\mathcal{A}} Q_h^t(x_h^t,a))\\ &= \left(Q_h^t - Q_h^\star\right)(x_h^t,a_h^t) + Q_h^\star(x_h^t,a_h^t) - V_h^{\pi_t}(x_h^t).\end{aligned}$$

According to Equation (8), if we let $m_h^t = m_h^t(x_h^t,a_h^t)$ and $t_1(x_h^t,a_h^t) < \cdots < t_{m_h^t}(x_h^t,a_h^t) < t$ be the episode indices where the pair (x_h^t,a_h^t) was encountered at step h before time t, then we have

$$\sum_{t=1}^n \left(Q_h^t - Q_h^\star\right)(x_h^t,a_h^t) = \sum_{t=1}^n \left(\alpha_{m_h^t}^0 H + \sum_{i=1}^{m_h^t} \alpha_{m_h^t}^i \phi_{h+1}^{t_i(x_h^t,a_h^t)} + \frac{3c\mathbf{1}\left\{m>0\right\}}{\sqrt{m_h^t}}\right).$$

We bound each of the three terms above separately. First recall $\alpha_{m_h^t}^0$ is 1 only if episode t is the first time we encounter the pair (x_h^t,a_h^t). Since there are at most $|\mathcal{X}||\mathcal{A}|$ pairs, we have $\sum_{t=1}^n \alpha_{m_h^t}^0 H \le |\mathcal{X}||\mathcal{A}|H$. For the second term, we regroup the summation by noting that for a fixed $t' = 1,\ldots n$, the part involving $\phi_{h+1}^{t'}$ is $\left(\sum_{t=m_h^{t'}+1}^{m_h^{n+1}-1} \alpha_t^{m_h^{t'}}\right)\phi_{h+1}^{t'}$ and thus

$$\sum_{t=1}^n\sum_{i=1}^{m_h^t} \alpha_{m_h^t}^i \phi_{h+1}^{t_i(x_h^t,a_h^t)} \le \sum_{t'=1}^n \left(\sum_{t=m_h^{t'}+1}^{\infty} \alpha_t^{m_h^{t'}}\right)\phi_{h+1}^{t'} \le \left(1+\frac{1}{H}\right)\sum_{t=1}^n \phi_{h+1}^t$$

where the last step again uses a property of the delicate choice of coefficient: $\sum_{t=i}^\infty \alpha_t^i = 1+\frac{1}{H}$. Finally, the third term can be bounded similarly to the UCB proof (Theorem 2 of Lecture 8):

$$\begin{aligned}\sum_{t=1}^n \frac{3c\mathbf{1}\left\{m_h^t>0\right\}}{\sqrt{m_h^t}} &= \sum_{x,a}\sum_{t=1}^n \frac{3c\mathbf{1}\left\{(x_h^t,a_h^t)=(x,a), m_h^t(x,a)>0\right\}}{\sqrt{m_h^t(x,a)}}\\ &\le 6c\sum_{x,a}\sqrt{m_h^n(x,a)} \le 6c\sqrt{n|\mathcal{X}||\mathcal{A}|},\end{aligned}$$

where the last step is by Cauchy-Schwarz inequality and the fact $\sum_{x,a} m_h^n(x,a) \le n$. Combining everything we have shown

$$\begin{aligned}\sum_{t=1}^n \delta_h^t &\le |\mathcal{X}||\mathcal{A}|H + \left(1+\frac{1}{H}\right)\sum_{t=1}^n \phi_{h+1}^t + 6c\sqrt{n|\mathcal{X}||\mathcal{A}|} + Q_h^\star(x_h^t,a_h^t) - V_h^{\pi_t}(x_h^t)\\ &= |\mathcal{X}||\mathcal{A}|H + \delta_{h+1}^t + \frac{1}{H}\sum_{t=1}^n \phi_{h+1}^t + 6c\sqrt{n|\mathcal{X}||\mathcal{A}|} + Q_h^\star(x_h^t,a_h^t) - V_{h+1}^\star(x_{h+1}^t)\\ &\quad + V_{h+1}^{\pi_t}(x_{h+1}^t) - V_h^{\pi_t}(x_h^t)\\ &\le |\mathcal{X}||\mathcal{A}|H + \left(1+\frac{1}{H}\right)\sum_{t=1}^n \delta_{h+1}^t + 6c\sqrt{n|\mathcal{X}||\mathcal{A}|} + \mathbb{P}_h V_{h+1}^\star(x_h^t,a_h^t) - V_{h+1}^\star(x_{h+1}^t)\\ &\quad + V_{h+1}^{\pi_t}(x_{h+1}^t) - \mathbb{P}_h V_{h+1}^{\pi_t}(x_h^t,a_h^t)\end{aligned}$$

where the last step uses the fact $\phi_h^t \le \delta_h^t$ since $V_h^{\pi_t}(x) \le V_h^\star(x)$ by the optimality of $V_h^\star$, and the Bellman equations (1) and (2). Applying the above connection between $\sum_{t=1}^n \delta_h^t$ and $\sum_{t=1}^n \delta_{h+1}^t$ repeatedly and the fact $\left(1+\frac{1}{H}\right)^H \le e$ and $c \approx \sqrt{H^3}$ we have

$$\begin{aligned}\sum_{t=1}^n \delta_1^t &\le \mathcal{O}\left(|\mathcal{X}||\mathcal{A}|H^2 + \sqrt{n|\mathcal{X}||\mathcal{A}|H^5}\right)\\ &\quad + \mathcal{O}\left(\sum_{t=1}^n\sum_{h=1}^H \mathbb{P}_h V_{h+1}^\star(x_h^t,a_h^t) - V_{h+1}^\star(x_{h+1}^t) + V_{h+1}^{\pi_t}(x_{h+1}^t) - \mathbb{P}_h V_{h+1}^{\pi_t}(x_h^t,a_h^t)\right).\end{aligned}$$

Finally, it remains to realize that the last summation is again the sum of a martingale difference sequence since $x_{h+1}^t \sim \mathbb{P}_h(\cdot|x_h^t,a_h^t)$, and therefore is of order $H\sqrt{nH}$ with high probability via Azuma inequality. This finishes the proof.

Remark. We point out that the dominating part of the regret bound is from the term $\sum_t 1/\sqrt{m_h^t}$, which again is similar to the key term in the proof of UCB (Theorem 2 of Lecture 8) and comes from the bonus term b_m. With a more careful design of the bonus term, the regret can be improved to $\mathcal{O}\left(\sqrt{n|\mathcal{X}||\mathcal{A}|H^4}\right)$. On the other hand, it can be shown that regret $\Omega\left(\sqrt{n|\mathcal{X}||\mathcal{A}|H^3}\right)$ is unavoidable for this problem, and closing the gap with model-free algorithms remains open. A matching upper bound can be achieved by a model-based algorithm though.

References

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.