Learning in Tabular MDPs: Upper Confidence Bound Value Iteration

Wen Sun

1 Preliminaries

We consider episodic finite horizon MDP with horizon H, $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \{r_h\}_{h=0}^{H-1}, \{P_h\}_{h=0}^{H-1}, H, s_0\}$, where s_0 is a fixed initial state, $r_h : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ and $P_h : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$ are time-dependent reward function and transition kernel. We assume $\mathcal{S}$ and $\mathcal{A}$ are finite and denote $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$. Note that for time-dependent finite horizon MDP, the optimal policy will be time-dependent as well. For simplicity, we overload notations a bit and denote $\pi = \{\pi_0, \dots, \pi_{H-1}\}$, where each $\pi_h : \mathcal{S} \mapsto \mathcal{A}$. We also denote $V^\pi := V_0^\pi(s_0)$, i.e., the expected total reward of π starting at $h = 0$ and s_0.

We define the learning protocol below. Learning happens in an episodic setting. Every episode k, learner first proposes a policy π^k based on all the history information up to the end of episode $k - 1$. The learner then executes π^k in the underlying MDP to generate a single trajectory $\tau^k = \{s_h^k, a_h^k\}_{h=0}^{H-1}$ with $a_h = \pi_h^k(s_h^k)$ and $s_{h+1}^k \sim P_h(\cdot | s_h^k, a_h^k)$. The goal of the learner is to minimize the following cumulative regret over N episodes:

$$\text{Regret} := \mathbb{E}\left[\sum_{k=1}^{N}\left(V^\star - V^{\pi^k}\right)\right],$$

where the expectation is with respect to the randomness of the MDP environment and potentially the randomness of the learner (i.e., the learner might make decisions in a randomized fashion).

2 Algorithm and Regret Bound

Algorithm 1 UCBVI

```
for $n = 1 \dots N$ do
    Compute $\widehat{P}_h^n$ for all $h$ (Eq. 1)
    Compute reward bonus $b_h^n$ for all h (Eq. 2)
    Run Value-Iteration on $\{\widehat{P}_h^n, r_h + b_h^n\}_{h=0}^{H-1}$ (Eq. 3)
    Set $\pi^n$ as the returned policy of VI.
end for
```

In this section, we present the UCBVI algorithm from Azar et al. (2017).

We first define some notations below. Consider the very beginning of episode n. We use the history information up to the end of episode $n - 1$ (denoted as $\mathcal{H}_{<n}$) to form some statistics. Specifically, we define:

$$N_h^n(s, a, s') = \sum_{i=1}^{n-1} \mathbf{1}\{(s_h^i, a_h^i, s_{h+1}^i) = (s, a, s')\}, \forall h, s, a, s',$$

$$N_h^n(s, a) = \sum_{i=1}^{n-1} \mathbf{1}\{(s_h^i, a_h^i) = (s, a)\}, \forall h, s, a.$$

Namely, we maintain counts of how many times s, a, s' and s, a are visited at time step h from the beginning of the learning process to the end of the episode $n-1$. We use these statistics to form an empirical model:

$$\widehat{P}_h^n(s'|s,a) = \frac{N_h^n(s,a,s')}{N_h^k(s,a)}, \forall h, s, a, s'. \tag{1}$$

We will also use the counts to define a *reward bonus*, denoted as $b_h(s,a)$ for all h, s, a. Denote $L := \ln\left(SAHN/\delta\right)$ (δ as usual represents the failure probability which we will define later). We define reward bonus as follows:

$$b_h^n(s,a) = H\sqrt{\frac{L}{N_h^n(s,a)}}. \tag{2}$$

With reward bonus and the empirical model, the learner uses *Value Iteration* on the empirical transition $\widehat{P}_h^n$ and the combined reward $r_h + b_h^n$. Starting at H (note that H is a fictitious extra step as an episode terminates at $H-1$), we perform dynamic programming all the way to $h = 0$:

$$\begin{aligned}
&\widehat{V}_H^n(s) = 0, \forall s,\\
&\widehat{Q}_h^n(s,a) = \min\left\{ r_h(s,a) + b_h^n(s,a) + \widehat{P}_h^n(\cdot|s,a)\cdot \widehat{V}_{h+1}^n,\ H \right\},\\
&\widehat{V}_h^n(s) = \max_a \widehat{Q}_h^n(s,a), \pi_h^n(s) = \operatorname*{argmax}_a \widehat{Q}_h^n(s,a), \forall h, s, a.
\end{aligned} \tag{3}$$

Note that when using $\widehat{V}_{h+1}^n$ to compute $\widehat{Q}_h^n$, we truncate the value by H. This is because we know that due to the assumption that $r(s,a) \in [0,1]$, no policy's Q value will ever be larger than H.

Denote $\pi^n = \{\pi_0^n, \ldots, \pi_{H-1}^n\}$. Learner then executes π^n in the MDP to get a new trajectory τ^n.

UCBVI repeats the above procedure for N episodes.

We will prove the following theorem.

Theorem 1 (Regret Bound of UCBVI). *UCBVI achives the following regret bound:*

$$\textit{Regret} := \mathbb{E}\left[\sum_{k=1}^{N}\left(V^\star - V^{\pi^k}\right)\right] \le 2H^2S\sqrt{AN\cdot\ln(SAH^2N^2)} = \widetilde{O}\left(H^2S\sqrt{AN}\right)$$

Remark While the above regret is sub-optimal, the algorithm we presented here indeed achieves a sharper bound in the leading term $\widetilde{O}(H^2\sqrt{SAN})$ (Azar et al., 2017), which gives the tight dependency bound on S, A, N. The dependency on H is not tight and tightening the dependency on H requires modifications to the reward bonus (use Bernstein rather than Hoeffding for reward bonus design).

3 Analysis

We prove the main theorem in this section.

We start with Hoeffding's inequality and union bound to bound state-action wise model error.

Lemma 2 (State-action wise ℓ_1 model error). *Fix $\delta \in (0,1)$. For all $n \in [1, \ldots, N], s \in \mathcal{S}, a \in \mathcal{A}, h \in [0, \ldots, H-1]$, with probability at least $1-\delta$, we have:*

$$\left\|\widehat{P}_h^n(\cdot|s,a) - P_h^\star(\cdot|s,a)\right\|_1 \le \sqrt{\frac{S\ln(SAHN/\delta)}{N_h^n(s,a)}}.$$

We have seen exactly the same bound in the generative model lecture.

The following lemma is still about model error, but this time we consider an average model error.

Lemma 3 (State-action wise average model error)**.** *Fix $\delta \in (0,1)$. For all $n \in [1, \dots, N], s \in \mathcal{S}, a \in \mathcal{A}, h \in [0, \dots, H-1]$, and any fixed $f : \mathcal{S} \to [0, H]$, with probability at least $1-\delta$, we have:*

$$\left| \widehat{P}^n_h(\cdot|s,a) \cdot f - P^\star_h(\cdot|s,a) \cdot f \right| \leq H \sqrt{\frac{\ln(SAHN/\delta)}{N^n_h(s,a)}}.$$

Please complete the proof of Lemma 3.

The above lemma is also used in the generative model lecture (recall we have $V^\star$ as f here). The key here is that f (and $V^\star$) needs to be independent of the data that used to form $\widehat{P}^n_h$. This obviously is true for $V^\star$ which is a deterministic quantity that exists at the moment when the MDP is specified.

We denote the two inequalities in Lemma 2 and Lemma 3 as event $\mathcal{E}_{model}$. Note that the failure probability of $\mathcal{E}_{model}$ is at most 2δ. Below we condition on $\mathcal{E}_{model}$ being true (we deal with failure event at the very end).

We also need the following simulation lemma (recall the simulation lemma we had in the generative model lecture here).

Lemma 4 (Simulation Lemma)**.** *Consider an arbitrary reward function $\hat{r}_h$ and an arbitrary transition kernel $\widehat{P}_h$ for $h \in [0, \dots, H-1]$. For any policy π, we have:*

$$V^\pi - \widehat{V}^\pi = \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d^\pi_h} \left[r_h(s,a) - \hat{r}_h(s,a) + \mathbb{E}_{s'\sim P^\star_h(\cdot|s,a)} \widehat{V}^\pi_{h+1}(s') - \mathbb{E}_{s'\sim \widehat{P}(\cdot|s,a)} \widehat{V}^\pi_{h+1}(s') \right]$$

Please complete the proof of the above simulation lemma.

One thing we want to emphasize here is that the above simulation lemma indeed applies to any MDPs (i.e., it has nothing to do with the discrete nature of $\mathcal{S}$ and $\mathcal{A}$), and it is a straight equality! Simulation Lemma is one of the most important lemmas one use over and over again in proving regret bounds for RL algorithms.

Now we study the effect of reward bonus. Similar to the idea in multi-armed bandits, we want to pick a policy π^n, such that the value of π^n in under the combined reward $r_h + b^n_h$ and the empirical model $\widehat{P}^n_h$ is optimistic, i.e., we want $\widehat{V}^n_0(s_0) \geq V^\star_0(s_0)$. The following lemma shows that via reward bonus, we are able to achieve this optimism.

Lemma 5 (Optimism)**.** *Assume $\mathcal{E}_{model}$ is true. For all episode n, we have:*

$$\widehat{V}^n_0(s_0) \geq V^\star_0(s_0),$$

where $\widehat{V}^n_h$ is computed based on VI in Eq. 3.

Proof. We prove via induction. Below we provide a proof sketch.

Starting at $h+1$, and assuming we have $\widehat{V}^k_{h+1}(s) \geq V^\star_{h+1}(s)$ for all s, we move to h below.

$$\begin{aligned}
\widehat{Q}^n_h(s,a) - Q^\star_h(s,a) &= b^n_h(s,a) + \widehat{P}^n_h(\cdot|s,a) \cdot \widehat{V}^n_{h+1} - P^\star_h(\cdot|s,a) \cdot V^\star_{h+1} \\
&\geq b^n_h(s,a) + \widehat{P}^n_h(\cdot|s,a) \cdot V^\star_{h+1} - P^\star_h(\cdot|s,a) \cdot V^\star_{h+1} \\
&= b^n_h(s,a) + \left(\widehat{P}^n_h(\cdot|s,a) - P^\star_h(\cdot|s,a) \right) \cdot V^\star_{h+1} \\
&\geq b^n_h(s,a) - H \sqrt{\frac{\ln(SAHN/\delta)}{N^n_h(s,a)}} \geq 0.
\end{aligned}$$

where the first inequality is from the inductive hypothesis, and the last inequality uses Lemma 3.

From $\widehat{Q}_{h+1}$, one can finish the proof by showing $\widehat{V}^n_h(s) \geq V^\star_h(s), \forall s$. □

Now we are almost ready to conduct the final steps.

Let us consider episode n and denote $\mathcal{H}_{<n}$ as the history up to the end of episode $n-1$. We consider bounding $V^\star - V^{\pi^n}$. Using Optimism and the simulation lemma, we can get the following result:

$$V^\star - V^{\pi^n} \le \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \le \sum_{h=0}^{H-1} \mathbb{E}_{s_h,a_h\sim d_h^{\pi^n}} \left[b_h^n(s_h,a_h) + \left(\widehat{P}_h^n(\cdot|s_h,a_h) - P^\star(\cdot|s_h,a_h) \right) \cdot \widehat{V}_{h+1}^{\pi^n} \right] \tag{4}$$

We prove the above two inequalities in the lecture. **Please provide a proof of the above inequality ((4)). Note that this is slightly different from the usual simulation lemma, as here we truncate $\widehat{V}$ by H during VI.**

We can bound $\left(\widehat{P}_h^n(\cdot|s_h,a_h) - P^\star(\cdot|s_h,a_h) \right) \cdot \widehat{V}_{h+1}^{\pi^n}$ using Lemma 2 with a Holder's inequality:

$$\begin{aligned} \left| \left(\widehat{P}_h^n(\cdot|s_h,a_h) - P^\star(\cdot|s_h,a_h) \right) \cdot \widehat{V}_{h+1}^{\pi^n} \right| &\le \left\| \widehat{P}_h^n(\cdot|s_h,a_h) - P^\star(\cdot|s_h,a_h) \right\|_1 \left\| \widehat{V}_{h+1}^{\pi^n} \right\|_\infty \\ &\le H\sqrt{\frac{S\ln(SANH/\delta)}{N_h^n(s,a)}}. \end{aligned}$$

Hence, back to per-episode regret $V^\star - V^{\pi^n}$, we get:

$$\begin{aligned} V^\star - V^{\pi^n} &\le \sum_{h=0}^{H-1} \mathbb{E}_{s_h,a_h\sim d_h^{\pi^n}} \left[b_h^n(s_h,a_h) + H\sqrt{S\ln(SAHN/\delta)/N_h^n(s_h,a_h)} \right] \\ &\le \sum_{h=0}^{H-1} \mathbb{E}_{s_h,a_h\sim d_h^{\pi^n}} \left[2H\sqrt{S\ln(SAHN/\delta)/N_h^n(s_h,a_h)} \right] \\ &= 2H\sqrt{\ln(SAHN/\delta)}\mathbb{E}\left[\sum_{h=0}^{H-1} \frac{1}{\sqrt{N_h^n(s_h^n,a_h^n)}} | \mathcal{H}_{<n} \right], \end{aligned}$$

where in the last term the expectation is taken with respect to the trajectory $\{s_h^n, a_h^n\}$ (which is generated from π^n) while conditioning on all history $\mathcal{H}_{<n}$ up to the end of episode $n-1$.

Now we sum all episodes together and take the failure event into consideration.

$$\begin{aligned} \mathbb{E}\left[\sum_{n=1}^N V^\star - V^{\pi^n} \right] &= \mathbb{E}\left[\mathbf{1}\{\mathcal{E}_{model}\} \left(\sum_{n=1}^N V^\star - V^{\pi^n} \right) \right] + \mathbb{E}\left[\mathbf{1}\{\overline{\mathcal{E}}_{model}\} \left(\sum_{n=1}^N V^\star - V^{\pi^n} \right) \right] \\ &\le \mathbb{E}\left[\mathbf{1}\{\mathcal{E}_{model}\} \left(\sum_{n=1}^N V^\star - V^{\pi^n} \right) \right] + 2\delta NH \\ &\le 2H\sqrt{S\ln(SAHN/\delta)}\mathbb{E}\left[\sum_{n=1}^N \sum_{h=0}^{H-1} \frac{1}{\sqrt{N_h^n(s_h^n,a_h^n)}} \right] + 2\delta NH \end{aligned}$$

We can bound the double summation term above using the following lemma:

Lemma 6. *Consider arbitrary N sequence of trajectories $\tau^n = \{s_h^n, a_h^n\}_{h=0}^{H-1}$ for $n = 1, \ldots, N$. We have*

$$\sum_{n=1}^N \sum_{h=0}^{H-1} \frac{1}{\sqrt{N_h^n(s_h^n,a_h^n)}} \le H\sqrt{SAN}.$$

We went through the proof in the lecture and **please complete the proof of the above lemma**
With this lemma, we can conclude that:

$$\mathbb{E}\left[\sum_{n=1}^{N} V^{\star} - V^{\pi^n}\right] \leq 2H^2 S\sqrt{AN\ln(SAHN/\delta)} + 2\delta NH.$$

Now set $\delta = 1/NH$, we get:

$$\mathbb{E}\left[\sum_{n=1}^{N} V^{\star} - V^{\pi^n}\right] \leq 2H^2 S\sqrt{AN\ln(SAH^2N^2)} + 2 = O\left(H^2 S\sqrt{AN\ln(SAH^2N^2)}.\right)$$

References

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. *arXiv preprint arXiv:1703.05449*, 2017.