7.3 Analysis

We will prove the following theorem in this section.

[Theorem 7.1] Regret Bound of UCBVI.

UCBVI achieves the following regret bound:

$$\text{Regret} := \mathbb{E}\left[\sum_{k=0}^{K-1}\left(V^*(s_0) - V^{\pi^k}(s_0)\right)\right]$$
$$\leq 20 H^2 S \sqrt{AK \cdot \ln(SAH^2K^2)} = \tilde{O}\left(H^2 S\sqrt{AK}\right)$$

Note: the dependency on H is not tight. By modifying the reward bonus using Bernstein inequlity, we could tighten it.

[Lemma 7.2] State-action wise model error.

Fix $\delta \in (0,1)$. For all $k \in \{0, \cdots, K-1\}$, $s \in \mathcal{S}$, $a \in \mathcal{A}$, $h \in \{0, \cdots, H-1\}$, with prob. $\geq 1-\delta$, we have that for any $f: \mathcal{S} \to [0, H]$:

$$\left|\left(\hat{P}_h^k(\cdot|s,a) - P_h(\cdot|s,a)\right)' f\right| \leq 8H\sqrt{S\ln(SAHK/\delta)/N_h^k(s,a)}$$

Pf: ① We consider a fixed tuple (s,a,k,h,f) first.

Recall the definition of $\hat{P}_h^k(s,a)$,

$$\hat{P}_h^k(\cdot|s,a)'f = \sum_{i=0}^{k-1} \mathbb{1}\{(s_h^i, a_h^i) = (s,a)\} f(s_{h+1}^i) / N_h^k(s,a)$$

Define $\mathcal{H}_{h,i}$ as the history from iter 0 to iter i including time step h.

Define $X_i = \mathbb{1}\{(s_h^i, a_h^i) = (s,a)\}\left[f(s_{h+1}^i) - \mathbb{E}_{s' \sim P_h(s,a)}[f(s')]\right]$,

We now show X_i is a martingale difference sequence.

(i) $\mathbb{E}[X_i | \mathcal{H}_{h,i}] = \mathbb{E}\left[f(s_{h+1}^i) - \mathbb{E}_{s'\sim P_h(s,a)}[f(s')] \,\middle|\, \mathcal{H}_{h,i}\right]$ or 0

$= 0$

Since $\mathbb{1}\{(s_h^i, a_h^i) = (s,a)\}$ is determined by $\mathcal{H}_{h,i}$

since $\mathbb{1}\{(s_h, a_h) = (s,a)\}$ is determined by $\mathcal{F}_{h,i}$.

(ii) We have $|X_i| = 0$ for $(s_h^i, a_h^i) \neq (s,a)$

$|X_i| \le H$ for $(s_h^i, a_h^i) = (s,a)$.

Then X_i is a martingale difference sequence. By Azuma-Hoeffding's inequality: with prob. $\geq 1-\delta$, we have:

$$\left|\sum_{i=0}^{k-1} X_i\right| = \left|\sum_{i=0}^{k-1} \mathbb{1}\{(s_h^i, a_h^i) = (s,a)\} f(s_{h+1}^i) - N_h^s(s,a) E_{s'\sim p_h(s,a)}[f(s')]\right|$$

$$\le 2H\sqrt{N_h^k(s,a)\ln(1/\delta)}$$

Apply union bound over $s \in S$, $a \in A$, $h \in \{0, \cdots, H-1\}$, $k \in \{0, \cdots, k-1\}$, with prob $\geq 1-\delta$,

$$\left|\sum_{i=0}^{k-1} \mathbb{1}\{(s_h^i, a_h^i) = (s,a) f(s_{h+1}^i) - N_h^s(s,a) E_{s'\sim p_h(s,a)}[f(s')]\right|$$

$$\le 2H\sqrt{N_h^k(s,a)\ln(SAKH/\delta)}$$

② next to cover all $f: S \to [0, H]$.

Note $\|f\|_2 \le H\sqrt{S}$ for all f, there exist a ε-net N_ε with $|N_\varepsilon| \le (1 + 2H\sqrt{S}/\varepsilon)^S$ s.t. for any $f \in [0, H]^S$, $\exists$ $f' \in N_\varepsilon$ s.t. $\|f - f'\|_2 \le \varepsilon$. [This is obvious since $[0,H]^S$ is obvious]

Then we have

$$\left|\sum_{i=0}^{k-1} \mathbb{1}\{(s_h^i, a_h^i) = (s,a)\} f(s_{h+1}^i) / N_h^k(s,a) - E_{s'\sim p_h(s,a)}[f(s')]\right|$$

$$\le \left|\sum_{i=0}^{k-1} \mathbb{1}\{(s_h^i, a_h^i) = (s,a)\} f'(s_{h+1}^i) / N_h^k(s,a) - E_{s'\sim p_h(s,a)}[f'(s')]\right|$$

$$+ \left|\sum_{i=0}^{k-1} \mathbb{1}\{(s_h^i, a_h^i) = (s,a)\} (f(s_{h+1}^i) - f'(s_{h+1}^i)) / N_h^k(s,a)\right|$$

$$+ \left|E_{s'\sim p_h(s,a)}(f(s') - f'(s'))\right|$$

$$\le 2H\sqrt{S\ln\left(SAKH(1+2H\sqrt{S}/\varepsilon^2)/\delta\right)/N_h^k(s,a)} + 2\varepsilon$$

Since ε^2-Net, $|f(s) - f'(s)|^2 \le \varepsilon^2 \Rightarrow |f(s) - f'(s)| \le \varepsilon$.

Now we set $\varepsilon^2 = 1/K$ and use the fact that $N_h^k(s,a) \le K$

we have:

$$\left| \sum_{i=0} \mathbb{1}\{(s_h^i, a_h^i) = (s,a)\} f(s_{h+1}^i) / N_h^k(s,a) - E_{s' \sim P_h(s,a)}[f(s')] \right|$$

$$\le 2H \sqrt{S \ln (SAKH(1+2HK\sqrt{S}/\delta)/N_h^k(s,a)} + 2/\sqrt{K}$$

$$\le 4H \sqrt{S \ln (4H^2 S^2 K^2 A/\delta)/N_h^k(s,a)}$$

$$\le 8H \sqrt{S \ln (HSKA/\delta)/N_h^k(s,a)}$$

which completes the proof of Lemma 7.2 □

[Lemma 7.3] State-action wise average model error under V^*.

Fix $\delta \in (0,1)$. For all $k \in \{0, \cdots, K-1\}$, with prob. $\geq 1-\delta$:

$$\left| \hat{P}_h^k(\cdot|s,a) \cdot V_{h+1}^* - P_h(\cdot|s,a) \cdot V_{h+1}^* \right| \le 2H \sqrt{\ln(SAHN/\delta)/N_h^k(s,a)}$$

Pf: Although we can bound the LHS by Lemma 7.2, since V^* is independent with data collected during learning, we can get a tighter upper bound.

① Consider a fixed tuple s, a, k, h first

$$\hat{P}_h^k(\cdot|s,a) V_{h+1}^* = \frac{1}{N_h^k(s,a)} \sum_{i=0}^{k-1} \mathbb{1}\{(s_h^i, a_h^i) = (s,a)\} V_{h+1}^*(s_{h+1}^i)$$

We define

$$X_i = \mathbb{1}\{(s_h^i, a_h^i) = (s,a)\} V_{h+1}^*(s_{h+1}^i) - E[\mathbb{1}\{(s_h^i, a_h^i) = (s,a)\} V_{h+1}^*(s_{h+1}^i) | H_{h,i}]$$

for $i = 0, \cdots, k-1$.

We have $E(X_i | H_{h,i}) = 0$, $|X_i| \le H$. Using Azuma-Hoeffding inequality, with prob. $\geq 1-\delta$

$$\left| \sum_{i=0}^{k-1} X_i \right| = \left| \sum_{i=0}^{k-1} \mathbb{1}\{(s_h^i, a_h^i) = (s,a)\} V_{h+1}^*(s_{h+1}^i) - N_h^k(s,a) E_{s' \sim P_h(s,a)} V_{h+1}^*(s') \right|$$

$$\le 2H \sqrt{N_h^k(s,a) \ln(1/\delta)}$$

Divide $N_h^k(s,a)$ on both side and use the fact $P_h^* V = E_{s \sim P_h}(V_{h+1}^*)$

we have:

$$\left| \hat{P}_h(\cdot|s,a)' V_{h+1}^* - P_h(\cdot|s,a)' V_{h+1}^* \right| \le 2H \sqrt{\ln(1/\delta)/N_h^k(s,a)}$$

with union bound over $S, A, [N], [H]$. we conclude the pf □

Now we condition on $\mathcal{E}_{model}$ (Lemma 7.2 & 7.3) being true.

We study the effect of reward bonus : We want π^k to be optimal under $r_h + b_h^k$ and empirical $\widehat{P}_h^k$, i.e. we want $\hat{V}_0^k(s_0) \geq V_0^*(s_0)$ for all s_0.

[Lemma 7.4] Optimism.

Assume $\mathcal{E}_{model}$ is true. For all episode k we have

$$\hat{V}_0^k(s_0) \geq V_0^*(s_0), \quad \forall s_0 \in S$$

where $\hat{V}_0^k(s_0)$ follows VI.

Pf: By induction.

(i) At time step H: $\hat{V}_H^k(s) = V_H^*(s) = 0$ for all s

(ii) Starting at $h+1$, assuming that $\hat{V}_{h+1}^k(s) \geq V_{h+1}^*(s)$ for $\forall s$

Consider any $s, a \in S \times A$. First, if $Q_h^k(s,a) = H$, then

$$Q_h^k(s,a) \geq Q_h^*(s,a)$$

$$\begin{aligned}
\hat{Q}_h^k(s,a) - Q_h^*(s,a) &= b_h^k(s,a) + \hat{P}_h^k(\cdot|s,a) \cdot \hat{V}_{h+1}^k - P_h^*(\cdot|s,a) V_{h+1}^* \\
&\geq b_h^k(s,a) + (\hat{P}_h^k(\cdot|s,a) - P_h^*(\cdot|s,a)) V_{h+1}^* \\
&\geq b_h^k(s,a) - 2H\sqrt{\frac{\ln(SAHK/\delta)}{N_h^k(s,a)}} \quad \text{(Lemma 7.3)} \\
&\geq 0
\end{aligned}$$

Then we have $V_h^k(s) \geq V_h^*(s) \quad \forall s$. $\square$

[Lemma 7.5] Consider $\tau^k = \{s_h^k, a_h^k\}_{h=0}^{H-1}$ for $k=0,\cdots,K-1$:

$$\sum_{k=0}^{K-1} \sum_{h=0}^{H-1} 1/\sqrt{N_h^k(s_h^k, a_h^k)} \leq 2H\sqrt{SAK}$$

Pf: LHS $= \displaystyle\sum_{h=0}^{H-1} \sum_{k=0}^{K-1} \frac{1}{\sqrt{N_h^k(s_h^k,a_h^k)}}$

$= \displaystyle\sum_{h=0}^{H-1} \sum \sum^{N_h^K(s,a)} \frac{1}{}$

$$= \sum_{h=0} \sum_{(s,a)} \sum_{i=1} \sqrt{i}$$

$$\leq 2 \sum_{h=0}^{H-1} \sum_{(s,a)} \sqrt{N_h^k(s,a)} \qquad \left(\frac{1}{\sqrt{i}} = \frac{2}{2\sqrt{i}} \leq \frac{2}{\sqrt{i}+\sqrt{i-1}} = 2(\sqrt{i}-\sqrt{i-1}) \right)$$

$$\leq \sum_{h=0}^{H-1} \sqrt{SA \sum_{s,a} N_h^k(s,a)} \qquad \text{Cauchy ineq.}$$

$$= H\sqrt{SAK} \qquad \square$$

Now we turn back to prove Th. 7.1:

$$\text{Regret} \leq 10H^2S\sqrt{AK\ln(SAH^2K^2)} = \tilde{O}(H^2S\sqrt{AK})$$

Pf [Th. 7.1]:

Consider episode k and history $\mathcal{H}_{<k}$.

① We bound $V^* - V^{\pi^k}$.

$$V_0^*(s_0) - V_0^{\pi^k}(s_0)$$

$$\leq \hat{V}_0^k(s_0) - V_0^{\pi^k}(s_0)$$

$$= \hat{Q}_0^k(s_0, \pi^k(s_0)) - Q_0^{\pi^k}(s_0, \pi^k(s_0))$$

$$\leq r_0(s_0,\pi^k(s_0)) + b_h^k(s_0,\pi^k(s_0)) + \hat{P}_0^k(\cdot|s_0,\pi^k(s_0)) \cdot \hat{V}_1^k \qquad \leq \text{ comes from } \hat{Q}_0^k \text{, which is } \min\{H, \sim\}.$$

$$-\left(r_0(s_0,\pi^k(s_0)) + P_0(\cdot|s_0,\pi^k(s_0)) \cdot V_1^{\pi^k}\right)$$

$$= b_h^k(s_0, \pi^h(s_0)) + \left(\hat{P}_0^k(\cdot|s_0,\pi^k(s_0)) - P_0(\cdot|s_0,\pi^k(s_0))\right) \cdot \hat{V}_1^k$$

$$+ P_0(\cdot|s_0,\pi^k(s_0)) \cdot \left(\hat{V}_1^k - V_1^{\pi^k}\right)$$

$$= \sum_{h=0}^{H-1} E_{s,a\sim d_h^{\pi^k}} \left[b_h^k(s,a) + \left(\hat{P}_h^k(\cdot|s,a) - P_h(\cdot|s,a)\right) \cdot \hat{V}_{h+1}^k \right] \quad (*)$$

With Lemma 7.2:

$$\text{RHS of } (*) \leq \sum_{h=0}^{H-1} E_{s,a\sim d_h^{\pi^k}} \left[b_h^k(s,a) + 8H\sqrt{S\ln(SAHK/\delta)/N_h^k(s,a)} \right]$$

$$\leq \sum_{h=0}^{H-1} E_{s_h,a_h\sim d_h^{\pi^k}} \left[10H\sqrt{SL/N_h^k(s,a)} \right]$$

$$= 10H\sqrt{SL}\; E\left[\sum_{h=0}^{H-1} 1/\sqrt{N_h^k(s,a)} \,\Big|\, \mathcal{H}_{<k} \right]$$

where $L = \ln(SAHK/\delta)$

② By summing all episodes together:

$$E\left[\sum_{k=0}^{K-1} V_0^*(s_0) - V_0^{\pi^k}(s_0) \right]$$

$$= E[\mathbb{1}\{\mathcal{E}_{model}\}(\sum_{k=0} V_0(s_0) - V_0(s_0))] + E[\mathbb{1}\{\mathcal{E}_{model}\}(\sum_{k=0} V(s_0) - V(s_0))]$$

$$\leq 10H\sqrt{SL}\, E\left[\sum_{k=0}^{K-1}\sum_{h=0}^{H-1} 1/\sqrt{N_h^k(s_h^k, a_h^k)}\right] + 28KH \quad (**)$$

With Lemma 7.5,

$$\text{RHS of } (**) \leq 10H\sqrt{SL} \cdot 2H\sqrt{SAK} + 28KH$$

$$= 20H^2S\sqrt{AK\ln(SAHK/\delta)} + 28KH$$

set $\delta = 1/KH$. we get

$$E\left[\sum_{k=0}^{K-1} V^*(s_0) - V^{\pi^k}(s_0)\right] \leq 20H^2S\sqrt{AK\ln(SAH^2K^2)} + 2$$

$$= O(H^2S\sqrt{AK\ln(SAH^2K^2)})$$

which completes the proof of Th. 7.1. $\square$