Lemma 1. *With probability at least $1-\delta$, for any $(h,k)\in[H]\times[K]$, $\mu^s\in\mathrm{BR}(\pi^s)$, and $\pi\in\Pi$*

$$L_h^{k-1}(\pi^*)-L_h^{k-1}(\pi)\le -2\sum_{s=1}^{k-1}\mathbb{E}_{\xi_h\sim\mu^s}[\ell_{\pi^s}(\pi;\xi_h)]+2\log(H|\Pi|/\delta).$$

Proof. Given $\pi\in\mathcal{H}$, we denote the random variable $X_{h,\pi}^k$ as

$$X_{h,\pi}^k=\log\left(\frac{\mathbb{P}_{h,\pi^*}(s_{h+1}^k|\,s_h^k,a_h^k)}{\mathbb{P}_{h,\pi}(s_{h+1}^k|\,s_h^k,a_h^k)}\right).$$

Now we define a filtration $\{\mathcal{F}_{h,k}\}_{k=1}^K$ as (B.25) in [1]. Thus we have $X_{h,\pi}^k\in\mathcal{F}_{h,k}$. Therefore, by applying Lemma D.1 in [1], we have that with probability at least $1-\delta$, for any $(h,k)\in[H]\times[K]$, and $\pi\in\Pi$, we have

$$-\frac{1}{2}\sum_{s=1}^{k-1}X_{h,\pi}^s\le\sum_{s=1}^{k-1}\log\mathbb{E}\left[\exp\left\{-\frac{1}{2}X_{h,\pi}^s\right\}|\,\mathcal{F}_{h,s-1}\right]+\log(H|\Pi|/\delta). \tag{1}$$

Meanwhile, by (B.27) in [1], for any $\mu^s\in\mathrm{BR}(\pi^s)$, the conditional expectation equals to

$$\mathbb{E}\left[\exp\left\{-\frac{1}{2}X_{h,\pi}^s\right\}|\,\mathcal{F}_{h,s-1}\right]=1-\mathbb{E}_{(s_h^s,a_h^s)\sim\mu^s}[D_H(\mathbb{P}_{h,\pi^*}(\cdot|\,s_h^s,a_h^s)||\mathbb{P}_{h,\pi}(\cdot|\,s_h^s,a_h^s))]. \tag{2}$$

Denote $\mathbb{E}_{(s_h^s,a_h^s)\sim\mu^s}[D_H(\mathbb{P}_{h,\pi^*}(\cdot|\,s_h^s,a_h^s)||\mathbb{P}_{h,\pi^s}(\cdot|\,s_h^s,a_h^s))]$ as $\mathbb{E}_{\xi_h\sim\mu^s}[\ell_{\pi^s}(\pi;\xi_h)]$. Using the fact $\log(x)\le x-1$ and substituting (2) into (1) finishes the proof. $\square$

Initializing a policy set $\mathcal{H}_\psi\leftarrow\mathcal{H}_{\mathrm{fin}}$, for all $k,\,l\in[K]$ with $k>l$, if $\pi^k\overset{\psi}{\sim}\pi^l$, we eliminate π^l from $\mathcal{H}_\psi$. The resulting $\mathcal{H}_\psi$ has the following property by its construction:

- $\mathcal{H}_\psi\subset\mathcal{H}_{\mathrm{fin}}$.
- $n^\psi(\mathcal{H}_\psi)\le n^\psi(\mathcal{H}_{\mathrm{fin}})$.

Lemma 2. *If for all $k\in[K]$ such that $\pi^k\in\mathcal{H}_\psi$, we have*

$$V(\psi(\pi^*),\pi^*)-V(\psi(\pi^k),\pi^k)\le c_k,$$

where $\{c_k\}_{k\in[K]}$ is a non-increasing sequence. Then, for all $k\in[K]$, we have

$$V(\psi(\pi^*),\pi^*)-V(\psi(\pi^k),\pi^k)\le c_k.$$

Proof. By definition, for all $k,\,l\in[K]$ with $k>l$ and $\pi^k\overset{\psi}{\sim}\pi^l$, we have

$$V(\psi(\pi^k),\pi^k)=V(\psi(\pi^l),\pi^l).$$

Thus,

$$V(\psi(\pi^*),\pi^*)-V(\psi(\pi^k),\pi^k)=V(\psi(\pi^*),\pi^*)-V(\psi(\pi^l),\pi^l).$$

Note that for all $k\in[K]$ such that $\pi^k\in\mathcal{H}_\psi$, we have

$$V(\psi(\pi^*),\pi^*)-V(\psi(\pi^k),\pi^k)\le c_k,$$

which implies $V(\psi(\pi^*),\pi^*) - V(\psi(\pi^l),\pi^l) \le c_k$. By the construction rule of $\mathcal{H}_\psi$, for all $l \in [K]$ with $\pi^l \notin \mathcal{H}_\psi$, we can always find a constant k' such that $k' > l$ and $\pi^{k'} \in \mathcal{H}_\psi$. Thus

$$V(\psi(\pi^*),\pi^*) - V(\psi(\pi^l),\pi^l) \le c_{k'} \le c_l. \qquad \square$$

Theorem 3. *Given an MDP with generalized eluder coefficient $d_{\mathrm{GEC}}(\cdot)$ and a finite hypothesis class $\mathcal{H}_{\mathrm{fin}}$ with $\pi^* \in \mathcal{H}_{\mathrm{fin}}$, by setting*

$$\eta = \sqrt{\frac{d_{\mathrm{GEC}}(1/\sqrt{HK})}{\log(Hn^\psi(\mathcal{H}_{\mathrm{fin}})/\delta)\cdot HK}},$$

the regret of the MEX algorithm applying on $\mathcal{H}_{\mathrm{fin}}$ with oracle ψ after K episodes is upper bounded by, with probability at least $1-\delta$,

$$\mathrm{Regret}(K) \lesssim \sqrt{d_{\mathrm{GEC}}(1/\sqrt{HK})\cdot \log(Hn^\psi(\mathcal{H}_{\mathrm{fin}})/\delta)\cdot HK}.$$

Proof. We decompose the regret into two terms,

$$\begin{aligned}\mathrm{Regret}(K) &\triangleq \sum_{k=1}^K V(\psi(\pi^*),\pi^*) - V(\psi(\pi^k),\pi^*)\\ &= \underbrace{\sum_{k=1}^K V(\psi(\pi^*),\pi^*) - V(\psi(\pi^k),\pi^k)}_{\text{Term (i)}} + \underbrace{\sum_{k=1}^K V(\psi(\pi^k),\pi^k) - V(\psi(\pi^k),\pi^*)}_{\text{Term (ii)}}.\end{aligned}$$

Term (i). By the choice of π^k, we have

$$V(\psi(\pi^*),\pi^*) - \eta\sum_{h=1}^H L_h^{k-1}(\pi^*) \le V(\psi(\pi^k),\pi^k) - \eta\sum_{h=1}^H L_h^{k-1}(\pi^k)$$

for all $k \in [K]$. Thus,

$$V(\psi(\pi^*),\pi^*) - V(\psi(\pi^k),\pi^k) \le \eta\sum_{h=1}^H (L_h^{k-1}(\pi^*) - L_h^{k-1}(\pi^k)). \tag{3}$$

Applying Lemma 1, we have that with probability at least $1-\delta$, for any $(h,k) \in [H]\times[K]$ and all $\pi \in \mathcal{H}_\psi$,

$$L_h^{k-1}(\pi^*) - L_h^{k-1}(\pi) \le -2\sum_{s=1}^{k-1} \mathbb{E}_{\xi_h\sim\psi(\pi^s)}[\ell_{\pi^s}(\pi;\xi_h)] + 2\log(H|\mathcal{H}_\psi|/\delta).$$

Substituting the above equation into (3) gives us, with probability at least $1-\delta$, for all $k \in [K]$ with $\pi^k \in \mathcal{H}_\psi$, we have

$$V(\psi(\pi^*),\pi^*) - V(\psi(\pi^k),\pi^k) \le -2\eta\sum_{h=1}^H\sum_{s=1}^{k-1} \mathbb{E}_{\xi_h\sim\psi(\pi^s)}[\ell_{\pi^s}(\pi^k;\xi_h)] + 2H\eta\log(H|\mathcal{H}_\psi|/\delta)$$

We define c_k as

$$c_k \triangleq -2\eta\sum_{h=1}^H\sum_{s=1}^{k-1} \mathbb{E}_{\xi_h\sim\psi(\pi^s)}[\ell_{\pi^s}(\pi^k;\xi_h)] + 2H\eta\log(H|\mathcal{H}_\psi|/\delta).$$

The sequnce $\{c_k\}_{k\in[K]}$ is a non-increasing sequence. Applying Lemma 2 gives us, with probability at least $1-\delta$, for all $k\in[K]$, we have

$$V(\psi(\pi^*),\pi^*)-V(\psi(\pi^k),\pi^k)\leq c_k.$$

Summing over $[K]$ gives us, with probability $1-\delta$,

$$\begin{aligned}\text{Term (i)}&\leq\sum_{k=1}^{K}c_k\\&=-2\eta\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s=1}^{k-1}\mathbb{E}_{\xi_h\sim\psi(\pi^s)}[\ell_{\pi^s}(\pi^k;\xi_h)]+2H\eta\log(H|\mathcal{H}_\psi|/\delta)\\&\leq-2\eta\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s=1}^{k-1}\mathbb{E}_{\xi_h\sim\psi(\pi^s)}[\ell_{\pi^s}(\pi^k;\xi_h)]+2H\eta\log(Hn_\psi(\mathcal{H}_{\text{fin}})/\delta).\end{aligned}$$

Term (ii). Follow the proof of Theorem 4.4 in [1],

$$\text{Term (ii)}\leq 2\eta\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s=1}^{k-1}\mathbb{E}_{\xi_h\sim\psi(\pi^s)}[\ell_{\pi^s}(\pi^k;\xi_h)]+\frac{d_{\text{GEC}}(\varepsilon_{\text{conf}})}{8\eta}+\sqrt{d_{\text{GEC}}(\varepsilon_{\text{conf}})HK}+\varepsilon_{\text{conf}}HK.$$

Combining Term (i) and Term (ii).

$$\begin{aligned}\text{Regret}(K)&=\text{Term (i)}+\text{Term (ii)}\\&\leq 2\eta KH\log(Hn^\psi(\mathcal{H}_{\text{fin}})/\delta)+\frac{d_{\text{GEC}}(\varepsilon_{\text{conf}})}{8\eta}+\sqrt{d_{\text{GEC}}(\varepsilon_{\text{conf}})HK}+\varepsilon_{\text{conf}}HK.\end{aligned}$$

Set $\varepsilon_{\text{conf}}=1/\sqrt{HK}$ and

$$\eta=\sqrt{\frac{d_{\text{GEC}}(1/\sqrt{HK})}{\log(Hn^\psi(\mathcal{H}_{\text{fin}})/\delta)\cdot HK}}$$

leads to the proof. □

Bibliography

[1] Zhihan Liu, Miao Lu, Wei Xiong, Han Zhong, Hao Hu, Shenao Zhang, Sirui Zheng, Zhuoran Yang, and Zhaoran Wang. One Objective to Rule Them All: A Maximization Objective Fusing Estimation and Planning for Exploration. may 2023.