Regret Analysis for MEX

by Jack Yansong Li
University of Illinois Chicago
Email: yli340@uic.edu

1 Regret Analysis for Finite Hypothesis Set

Through this note, we denote π^k as the guesses policy of player 2 in episode k and π^* as the true policy of player 2. For each player 2's policy π, the set of all best response policies is denoted as $\mathrm{BR}(\pi)$, i.e.,

$$\mathrm{BR}(\pi) = \operatorname*{argmax}_{\mu \in \mathcal{U}} V(\mu, \pi),$$

where $\mathcal{U}$ is the set of all possible policies for player 1. The hypothesis set of all possible policies of player 2 is denoted as $\mathcal{H}$.

1.1 Oracle and Type Number

For any player 2's policy π, we assume the existence of an oracle that can return a best response from $\mathrm{BR}(\pi)$.

Definition 1. ***(Oracle)*** *A best response oracle ψ refers to a function that, upon receiving policies as input, yields a best response as its output, i.e., ψ is a function $\psi: \mathcal{H} \rightarrow \mathcal{U}$ such that*

$$\psi(\pi) \in \mathrm{BR}(\pi).$$

With the definition of an oracle, we can categorize policies within a hypothesis set into various types.

Definition 2. ***(ψ-type)*** *We call two policies π and π' to be of the same type under oracle ψ if for any $\mu = \psi(\pi)$ and $\mu' = \psi(\pi')$ we have*

$$V(\mu, \pi) = V(\mu', \pi').$$

The relationship is denoted as $\pi \overset{\psi}{\sim} \pi'$. On the contrary, two policies π and π' not of the same type under oracle ψ is denoted as $\pi \overset{\psi}{\nsim} \pi'$.

Definition 3. *We call a set of policies Π be type-independent under oracle ψ if for all $\pi \in \Pi$ and $\pi' \in \Pi$ such that $\pi \neq \pi'$, we have $\pi \overset{\psi}{\nsim} \pi'$.*

The ψ-type characterization gives rise to a measurement of quantity for the set of policies $\mathcal{H}$, denoted by $n^\psi(\mathcal{H})$.

Definition 4. *Given a hypothesis set $\mathcal{H}$, the type number $n^\psi(\mathcal{H})$ under oracle ψ is defined as the size of a largest type-independent subset of $\mathcal{H}$, i.e.,*

$$n^\psi(\mathcal{H}) = \max |\Pi|,$$

where $\Pi \subset \mathcal{H}$ and Π is type-independent under oracle ψ.

1.2 Regret Analysis

In this subsection, we restrict our discussion to cases where the cardinality of the hypothesis set is finite, i.e., $|\mathcal{H}|<\infty$. This condition is emphasized through the notation $\mathcal{H}_{\text{fin}}$. We also assume that the realization assumption holds, i.e., $\pi^* \in \mathcal{H}_{\text{fin}}$.

Theorem 5. *Given an MDP with generalized eluder coefficient $d_{\text{GEC}}(\cdot)$ and a finite hypothesis class $\mathcal{H}_{\text{fin}}$ with $\pi^* \in \mathcal{H}_{\text{fin}}$, by setting*

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

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

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

Proof. See Appendix 3.2 □

The sole term pertaining to the size of the hypothesis set is $n^{\psi}(\mathcal{H}_{\text{fin}})$. Consequently, the magnitude of regret is solely influenced by the type number associated with a hypothesis set, as opposed to the cardinality of the hypothesis set. This phenomenon occurs because policies that are categorized under the same type by policy ψ yield identical rewards when implemented in the MEX algorithm.

The type number $n^{\psi}(\mathcal{H}_{\text{fin}})$ depends on the choice of the oracle ψ, which makes it hard to verify when the explicit form of ψ is not given. However, we can introduce a stronger notion of type and verify the upper bound of $n^{\psi}(\mathcal{H}_{\text{fin}})$.

Definition 6. *(Strong) We call two policies π and π' to be of the same s-type if*

$$V(\mu,\pi) = V(\mu,\pi') = V(\mu',\pi) = V(\mu',\pi')$$

for all $\mu \in \text{BR}(\pi)$ and $\mu' \in \text{BR}(\pi')$. The relationship are denoted as $\pi \overset{s}{\sim} \pi'$.

Similar to the definition of type number under oracle ψ, we can define strong type number $n_{\text{stype}}(\mathcal{H})$.

Lemma 7. *$n^{\psi}(\mathcal{H}) \le n_{\text{stype}}(\mathcal{H})$ for all ψ be a best response oracle.*

2 Regret Analysis for Infinite Hypothesis Set

In this subsection, we discuss the cases where the cardinality of the hypothesis set is infinite, i.e., $|\mathcal{H}|=\infty$. This condition is emphasized through the notation $\mathcal{H}_{\text{inf}}$. We keep assume that the realization assumption holds, i.e., $\pi^* \in \mathcal{H}_{\text{inf}}$.

2.1 Approximate an Infinite Hypothesis Set by a Finite Hypothesis Set

A direct approach to handling an infinite hypothesis set is to approximate it as a finite hypothesis set. First, we outline what makes a good approximation.

Definition 8. *(ε_ψ-optimal approximation) A finite hypothesis set $\mathcal{H}_{\text{fin}}$ is called an ε_ψ-optimal approximation of $\mathcal{H}_{\text{inf}}$ if*

$$\min_{\pi\in\mathcal{H}} |V(\psi(\pi),\pi) - V(\psi(\pi^*),\pi^*)| \le \varepsilon_\psi$$

We denote π^*_{det} as ε_ψ-optimal policy defined as

$$\pi^*_{\text{det}} \triangleq \operatorname*{argmin}_{\pi \in \mathcal{H}} |V(\psi(\pi), \pi) - V(\psi(\pi^*), \pi^*)|.$$

In the following, we list some examples of the ε_ψ-optimal approximation. These examples are established based on the following lemma.

Lemma 9. *For any best response oracle ψ, if $\|\pi - \pi^*\| \leq \varepsilon$, then*

$$|V(\psi(\pi), \pi) - V(\psi(\pi^*), \pi^*)| \leq L_\psi \varepsilon,$$

where $L_\psi > 0$ is a constant.

Example 10. Given a finite hypothesis set $\mathcal{H}_{\text{fin}}$. For $\varepsilon > 0$, Define an infinite hypothesis $\mathcal{H}_{\text{inf}}$ as

$$\mathcal{H}_{\text{inf}} \triangleq \{\pi | \, \|\pi - \pi'\| \leq \varepsilon, \pi' \in \mathcal{H}_{\text{fin}}\}.$$

For the constructed $\mathcal{H}_{\text{inf}}$, $\mathcal{H}_{\text{fin}}$ is an $L_\psi \varepsilon$-optimal approximation of $\mathcal{H}_{\text{inf}}$.

Example 11. Given a specific neural network structure $\mathcal{N}$, we define $\mathcal{H}_{\text{inf}}$ as the set comprising all neural networks characterized by the set all possible parameters Θ that is in accordance with the specified structure, formally represented as,

$$\mathcal{H}_{\text{inf}} = \{\pi | \, \pi \in \mathcal{N}(\theta), \theta \in \Theta\}.$$

We proceed to create a discretization of Θ, denoted as $\hat{\Theta}$. The finite approximation set $\mathcal{H}_{\text{fin}}$ is defined as

$$\mathcal{H}_{\text{fin}} = \{\pi | \, \pi \in \mathcal{N}(\theta), \theta \in \hat{\Theta}\}.$$

By the choice of discretization interval, we can ensure that $\|\pi - \pi^*\| \leq \varepsilon$. Consequently, the discretized set $\mathcal{H}_{\text{fin}}$ serves as $L_\psi \varepsilon$-optimal approximation of $\mathcal{H}_{\text{inf}}$.

2.2 Regret Analysis

Now, given an infinite hypothesis set $\mathcal{H}_{\text{inf}}$ with an ε_ψ-optimal approximation set $\mathcal{H}_{\text{fin}}$, we are prepared to execute the MEX algorithm within the confines of $\mathcal{H}_{\text{fin}}$. The regret analysis is given in the following Theorem.

Theorem 12. *Given an MDP with generalized eluder coefficient $d_{\text{GEC}}(\cdot)$ and an infinite hypothesis class $\mathcal{H}_{\text{inf}}$ with $\pi^* \in \mathcal{H}_{\text{inf}}$. For any ε_ψ-optimal approximation $\mathcal{H}_{\text{fin}}$ of $\mathcal{H}_{\text{inf}}$, by setting*

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

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

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

Proof. By the choice of π^k, we have

$$V(\psi(\pi^*_{\text{det}}), \pi^*_{\text{det}}) - \eta \sum_{h=1}^{H} L_h^{k-1}(\pi^*_{\text{det}}) \leq V(\mu^k, \pi^k) - \eta \sum_{h=1}^{H} L_h^{k-1}(\pi^k)$$

for all $k \in [K]$. By Definition 8,

$$V(\psi(\pi^*_{\text{det}}), \pi^*_{\text{det}}) \geq V(\psi(\pi^*), \pi^*) - \varepsilon_\psi.$$

Thus,

$$V(\psi(\pi^*), \pi^*) - V(\mu^k, \pi^k) \leq \eta \sum_{h=1}^{H} L_h^{k-1}(\pi^*_{\text{det}}) - \eta \sum_{h=1}^{H} L_h^{k-1}(\pi^k) + \varepsilon_\psi.$$

Follow the same procedure in the proof of Theorem 5 leads to the proof. □

Remark 13. The linear term $K\varepsilon_\psi$ cannot be eliminated. Consider the best case where $\pi^k = \pi^*_{\text{det}}$ for all $k \in [K]$. The regret is

$$\begin{aligned}\text{Regret}(K) &= \sum_{k=1}^{K} V(\psi(\pi^*_{\text{det}}), \pi^*_{\text{det}}) - V(\psi(\pi^*), \pi^*) \\ &= K(V(\psi(\pi^*_{\text{det}}), \pi^*_{\text{det}}) - V(\psi(\pi^*), \pi^*)) \\ &\leq K\varepsilon_\psi.\end{aligned}$$

3 Appendix

3.1 Lemmas

Lemma 14. *With probability at least $1-\delta$, for any $(h,k) \in [H] \times [K]$, $\mu^s \in \text{BR}(\pi^s)$, and $\pi \in \Pi$*

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

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

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

Meanwhile, by (B.27) in [1], for any $\mu^s \in \text{BR}(\pi^s)$, 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 $\log(x)\leq x-1$ and substituting (2) into (1) finishes the proof. □

3.2 Proof of Theorem 5

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

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

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

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

for any $\pi^k \overset{\psi}{\sim} \pi^{k\prime}$, we have

$$V(\psi(\pi^k),\pi^k) = V(\psi(\pi^{k\prime}),\pi^{k\prime}).$$

Thus, an upper bound for $V(\psi(\pi^*),\pi^*) - V(\mu^k,\pi^k)$ is also an upper bound for $V(\psi(\pi^*),\pi^*) - V(\mu^{k\prime},\pi^{k\prime})$. Applying Lemma 14, we have that with probability at least $1-\delta$, for any $(h,k)\in[H]\times[K]$, $\mu^s=\psi(\pi^s)$ and $\pi^k\in\mathcal{H}_{\mathrm{fin}}$,

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

Substituting the above equation into (3) gives us that with probability at least $1-\delta$, for any $k\in[K]$, $\mu^s=\psi(\pi^s)$ and $\pi^k\in\mathcal{H}_{\mathrm{fin}}$

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

Summing over $[K]$ gives us

$$\text{Term (i)} \leq -2\eta\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s=1}^{k-1} \mathbb{E}_{\xi_h\sim\mu^s}[\ell_{\pi^s}(\pi;\xi_h)] + 2\eta KH\log(Hn^{\psi}(\mathcal{H}_{\mathrm{fin}})/\delta).$$

Term (ii). Follow the proof of Theorem 4.4 in [1], we have that for all $\mu^s=\psi(\pi^s)$

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

Combining Term (i) and Term (ii).

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

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.