Case Study: Realizability

Notation of hypothesis sets:

- Set of all potential partners: $\mathcal{H}^*$.
- Hypothesis set of partners: $\mathcal{H}$.

1 Summary

Question 1. *(SH)* Could you summarize the main differences between this note and the one you sent me last week?

- Fact: GEC depends on ε and the choice of ε usually depends on K and H (time horizon, in MAB example, $H=1$).
- Fact: The regret bound is still linear if $d(\varepsilon)$ grows linear w.r.t. K.
- We showed that: By choosing $\varepsilon = |\mathcal{H}|/K$, the GEC $d(\varepsilon)=0$. In this case, the regret is upper bounded by $\varepsilon K = |\mathcal{H}|$.

2 Generalized Eluder Dimension

Consider a normal-form game defined as $(2, A_{\text{joint}} = [N]^2, r_{\text{joint}} = (V, V))$. The shared reward function V is defined as

$$V(a,b) = \begin{cases} 1 & \text{if} \quad a = b \\ 0 & \text{if} \quad a \neq b \end{cases}.$$

In this game, player 2 only adopts a pure strategy, i.e., $\mathcal{H}^* = [N]$. The game is played for K episodes, given the partner's real pure strategy π^*.

In episode k, the AI agent chooses a response $\psi(\pi^k)$, where ψ is a best response oracle, and

$$\pi^k \in \arg\max_{\pi \in \mathcal{H}} V(\psi(\pi), \pi) - \eta L^k(\pi). \tag{1}$$

The estimation loss L^k at episode k is defined as

$$L^k = \sum_{s=0}^{k-1} L(\pi, \psi(\pi^s), r^s).$$

where $L^0(\pi) \equiv 0$, r^{k-1} is the reward of episode $k-1$, and L is defined as

$$L(\pi, a, r) = \begin{cases} 0 & \text{if } V(a,\pi) = r \\ 1 & \text{if } V(a,\pi) \neq r \end{cases}.$$

The regret of the game is defined as

$$\sum_{k=1}^{K} V(\psi(\pi^*), \pi^*) - V(\psi(\pi^k), \pi^*),$$

which can be decomposed into the sum of

$$\sum_{k=1}^{K} V(\psi(\pi^*), \pi^*) - V(\psi(\pi^k), \pi^k) \tag{2}$$

and

$$\sum_{k=1}^{K} V(\psi(\pi^k), \pi^k) - V(\psi(\pi^k), \pi^*). \tag{3}$$

In the following, we call (2) the *value guessed error* and (3) the *prediction error*. We have already proved the following result. The name of GEC refers to the generalized eluder coefficient, defined as

Definition 1. *(GEC)* Given $\varepsilon > 0$, a best response oracle ψ[1] there exist $d > 0$, such that for any sequence of player 2's pure strategy $\{\pi^k\}_{k \in [K]} \subset \mathcal{H}^*$, and $\{\psi(\pi^k)_{k \in [K]}\}$,

$$\sum_{k \in [K]} V(\psi(\pi^k), \pi^k) - V(\psi(\pi^k), \pi^*) \leq \left(d \sum_{k=1}^{K} \sum_{s=1}^{k-1} \ell(\pi^k; \pi^*) \right)^{1/2} + \sqrt{dK} + \varepsilon K, \tag{4}$$

where $\ell(\pi^k; \pi^*) = D_H(V(\psi(\pi^k), \pi^k), V(\psi(\pi^k), \pi^*))$. The smallest d that satisfies the above inequality is called GEC.

Intuitively, the existence of the GEC states that if the training error $\sum_{k=1}^{K} \sum_{s=1}^{k-1} \ell(\pi^k; \pi^*)$ is small, then the estimation error $\sum_{k \in [K]} V(\psi(\pi^k), \pi^k) - V(\psi(\pi^k), \pi^*)$ is also small.

In this note, we only consider the realizable case, i.e., $\pi^* \in \mathcal{H}$. The goal of this note is to show that in this game

- The GEC assumption holds with a GEC $d = 0$.
- The regret analysis adopted by the MEX paper [1] and the GEC paper [2] leads to a regret upper bound $|\mathcal{H}|$.

3 Regret analysis

Recall that the regret can be decomposed into

$$\underbrace{\sum_{k=1}^{K} V(\psi(\pi^*), \pi^*) - V(\psi(\pi^k), \pi^k)}_{\text{value guessed error}} + \underbrace{\sum_{k=1}^{K} V(\psi(\pi^k), \pi^k) - V(\psi(\pi^k), \pi^*)}_{\text{prediction error}}.$$

By (1), the type error is bounded by

$$\sum_{k} V(\psi(\pi^*), \pi^*) - V(\psi(\pi^k), \pi^k) \leq \eta \sum_{k \in [K]} L^k(\pi^*) - L^k(\pi^k) = -\eta \sum_{k \in [K]} L^k(\pi^k).$$

1. The oracle ψ can be non-best response in the original definition of GEC. The truth is, the definition of GEC depends on the choice of ψ and the discrepancy function ℓ.

In the original MEX paper, they showed that the term $-\sum_{k\in[K]} L^k(\pi^k)$ is upper bound by the negative of the training error, which cancels out the training error in the upper bound of the GEC term. The fact that $-\sum_{k\in[K]} L^k(\pi^k)$ is upper bound by $-\sum_{k=1}^{K}\sum_{s=1}^{k-1} \ell(\pi^k;\pi^*)$ means that a large training error implies a large $\sum_{k\in[K]} L^k(\pi^k)$. This is intuitive since both $L^k(\pi^k)$ and $\ell(\pi^k;\pi^*)$ serve as a measure that measures the distance between the guess partner's policy π^k and the true partner policy π^*. Now, back to our case, the term $-\sum_{k\in[K]} L^k(\pi^k)$ is

$$-\sum_{k\in[K]} L^k(\pi^k) = -\sum_{k\in K}\sum_{s=1}^{k} L(\pi^k, a^s, r^s) = 0.$$

Thus, the regret is entirely upper bounded by the prediction error, i.e.,

$$\mathrm{Reg}(K) = \underbrace{\sum_{k=1}^{K} V(\psi(\pi^k), \pi^k) - V(\psi(\pi^k), \pi^*)}_{\text{prediction error}}.$$

With the GEC, the GEC term can be upper bounded by the training error. Recall that the training error is

$$\sum_{k=1}^{K}\sum_{s=1}^{k-1} \ell(\pi^k;\pi^*) = \sum_{k=1}^{K} \frac{n(k-1)}{\sqrt{2}},$$

where the Hellinger distance $\ell(\pi^k;\pi^*)$ always equals $1/\sqrt{2}$ for all $k\in[K]$ where $\pi^k\neq\pi^*$ and $n(K)\triangleq|\{k|\, k\in[K], \pi^k\neq\pi^*\}|$. It is easy to verify that

$$n(k)\leq|\mathcal{H}|$$

in our setting. Thus, the game we created satisfies the *optimism* and the *small in-sample training error* property mentioned on [2] page 14. Following the same analysis as in [2] page 14, we have

$$\mathrm{Reg}(K)\leq\sqrt{dK|\mathcal{H}|}+\sqrt{dK}+\varepsilon K. \tag{5}$$

To achieve a sublinear regret, we need a d that grows sublinear w.r.t. K.

Remark 2. The GEC depends on ε, which further depends on K (See the discussion of burn-in cost after Definition 3.4 in [2]). Thus, the GEC also depends on K.

4 The existence of GEC

First to note that, we have

$$\left(d\sum_{k=1}^{K}\sum_{s=1}^{k-1} \ell(\pi^k;\pi^*)\right)^{1/2} = \left(d\sum_{k=1}^{K} \frac{n(k-1)}{\sqrt{2}}\right)^{1/2}$$

Also,

$$\sum_{k\in[K]} V(\psi(\pi^k), \pi^k) - V(\psi(\pi^k), \pi^*) = K-(K-n(K)).$$

Thus, for every $\varepsilon > 0$, if we can find d such that

$$n(K) \leq \left(d \sum_{k=1}^{K} \frac{n(k-1)}{\sqrt{2}} \right)^{1/2} + \sqrt{dK} + K,$$

then the GEC assumption holds. The above inequality is equivalent to

$$d \geq \left(\frac{n(K) - \varepsilon K}{\sqrt{\sum_{k=1}^{K} \frac{n(k-1)}{\sqrt{2}}} + \sqrt{K}} \right)^2.$$

Recall that $n(K) \leq |\mathcal{H}|$. Thus, an upper bound for GEC is

$$d(\varepsilon) = \left(\frac{|\mathcal{H}| - \varepsilon K}{\sqrt{\sum_{k=1}^{K} \frac{n(k-1)}{\sqrt{2}}} + \sqrt{K}} \right)^2.$$

Choose $\varepsilon = |\mathcal{H}| / K$ implies $d(\varepsilon) = 0$. Substituting into (5) gives

$$\mathrm{Reg}(K) \leq |\mathcal{H}|.$$

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.

[2] Han Zhong, Wei Xiong, Sirui Zheng, Liwei Wang, Zhaoran Wang, Zhuoran Yang, and Tong Zhang. GEC: A Unified Framework for Interactive Decision Making in MDP, POMDP, and Beyond. jun 2023.