Week 9 – Lecture 40

$$P(A) + Q(A^c) \geq \frac{1}{2} \exp\left(-D(P \| Q)\right)$$

Let $\nu = (P_i)_{i \in [K]}$

π - Policy

Let $P_{\nu\pi} \sim (A_1 X_1, A_2 X_2, A_3 X_3 \dots A_T, X_T)$

Also

$a_1, a_2 \dots a_t$ are discreet while $x_1, x_2 \dots x_t$ come from continous distribution

$$P_{\nu\pi}(a_1, x_1, a_2, x_2 \dots a_T, x_T) =$$

$$\overset{||}{P_\nu} \qquad \prod_{t=1}^{T} \underbrace{\pi(a_t | a_1, x_1, \dots a_{t-1}, x_{t-1})}_{\text{Policy}} \underbrace{P_{a_t}(x_t)}_{\text{reward from env. for action } a_t} \quad \text{— (1)}$$

Similarly: (diff. env. same policy) $\nu' = (P_i')_{i \in [K]}$

$$P_{\nu'\pi}(a_1, x_1, a_2, x_2, \dots a_t, x_t)$$

$$\overset{||}{P_{\nu'}} = \prod_{t=1}^{t} \pi(a_t | a_1, x_1 \dots a_{t-1}, x_{t-1}) P'_{a_t}(x_t) \quad \text{— (2)}$$

Lemma:

$$D(P_v \| P_{v'}) = \sum_{i=1}^{K} \mathbb{E}_v[N_i(T)] \, D(P_i \| P_i')$$

<u>Proof</u> :

$$\log \frac{dP_v}{dP_{v'}} (a_1 x_1 a_2 x_2 \ldots a_t x_t)$$

Putting ① & ② and simplifying

$$= \sum_{t=1}^{T} \log \frac{P_{a_t}(x_t)}{P'_{a_t}(x_t)}$$ ← This was for one realization

$$\mathbb{E}_v \left[\log \frac{dP_v}{dP_{v'}} (A_1 X_1 \ldots A_T X_T) \right] = D(P_v \| P_{v'})$$

*

$$= \mathbb{E}_v \left[\sum_{t=1}^{T} \log \frac{P_{A_T}(X_t)}{P'_{A_t}(X_t)} \right]$$

$$\boxed{* \quad D(P \| Q) = \int \log \frac{dP}{dQ}(\omega) \, dP(\omega)}$$

$$= \mathbb{E}_v \left[\sum_{t=1}^{T} \mathbb{E}_v \left[\log \frac{P_{A_t}(X_t)}{P'_{A_t}(X_t)} \middle| A_t \right] \right]$$ ← conditioned on A_t

$$= \mathbb{E}_v \left[\sum^{T} D(P_{A_t} \| P'_{A_t}) \right]$$ ← Now the

$$\Big]$$

expectation is only over the randomness of arm pulls

$$= \sum_{k=1}^{K} \mathbb{E}_{\nu}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t = k\} D(P_{A_t} \| P'_{A_t})\right]$$

$$= \sum_{k=1}^{K} \mathbb{E}_{\nu}[N_k(T)] \, D(P_{A_t} \| P'_{A_t})$$

Hence proved

Divergence between the two environments can be decomposed into divergence between arms

Theorem: $K > 1$, $T \geq K - 1$

$\forall \pi \; \exists$ a mean vector $\mu \in [0, 1]^K$ s.t.

$$R_T(\pi, \nu_\mu) \geq \frac{1}{27}\sqrt{(K-1)T}$$

[$\nu_\mu \rightarrow$ env. whose rewards come from μ]

$$\Rightarrow R^*_T(\mathcal{E}^K) \geq \frac{1}{27}\sqrt{(K-1)T}$$

minmax regret