E1 245: Online Prediction and Learning **Fall 2015**

Lecture 21 — October 15

Lecturer: Aditya Gopalan *Scribe: Sindhu P R*

21.1 A Minimax Lower Bound on Regret for Bandit Algorithms

We consider the multi-armed bandit problem with N actions. The reward of action i at time t is denoted as $g(i,t)$ which is either 0 or 1. An algorithm's choice of an action at time step t is denoted as I_t, to which the environment assigns the reward $g(I_t,t)$. The regret of an algorithm for T time steps, with respect to the best performing action is given as:

$$\max_{i\in[N]} \sum_{t=1}^{T} g(i,t) - \sum_{t=1}^{T} g(I_t,t).$$

Assuming the rewards $g(.,t)$, $\forall t$ are stochastic and in particular distributed according to the Bernoulli distribution, we have the following lower bound [1]:

Theorem 21.1. *Let the reward distributions $g(i,1), g(i,2), \ldots, g(i,T)$ be i.i.d Bernoulli random variables for all $i \in [N]$. Let sup be the supremum over all reward distributions and inf be the infimum over all algorithms (forecasters)*

$$\inf \sup \left[\max_{i\in[N]} \mathbb{E} \left[\sum_{t=1}^{T} g(i,t) - \sum_{t=1}^{T} g(I_t,t) \right] \right] \geq c_1 \sqrt{NT},$$

where the expectations are with respect to random generation of rewards and the algorithm's choice of actions.

There are two sources of randomness - the reward sequence and the algorithm's choice of actions. Once a gain sequence has been set for all actions, then the sequence $\{I_t\}$ results in one realisation of the reward sequence.

For the proof, we use the idea of stochastic Bandit models. Let $\varepsilon \in (0,1)$. For each $i \in [N]$, let the Bandit model i be the stochastic reward generating distribution, where all action's rewards are i.i.d Bernoulli($\frac{1-\varepsilon}{2}$) except for action i, whose rewards are i.i.d Bernoulli($\frac{1+\varepsilon}{2}$). Hence based on this definition, bandit model 0 corresponds to all actions having same reward distribution Bernoulli($\frac{1-\varepsilon}{2}$). Let $\mathbb{E}_i[.]$ denote the expectation for bandit model i. Then, $\forall i$ the following holds:

Lemma 21.2.

$$\max_{i\in[N]} \mathbb{E}_i \left[\sum_{t=1}^{T} g(i,t) - \sum_{t=1}^{T} g(I_t,t) \right] \geq T\varepsilon \left(1 - \frac{1}{N} - \sqrt{\frac{\varepsilon T}{2N} \log\left(\frac{1+\varepsilon}{1-\varepsilon} \right)} \right).$$

Consider deterministic bandits algorithm. Let $S_{i,T} = \sum_{t=1}^{T} \mathbb{1}_{\{I_t=i\}}$, $\forall i \in [N]$. This keeps track of the number of times action i was chosen by the algorithm. The proof for this lemma requires the following tools:

Let P, Q be two probability distributions over $\{1,2,\ldots,M\}$.

Definition 1. *The Total Variation distance [2] between P and Q is*

$$d_{TV}(P,Q) = \max_{S \subseteq [M]} (P[S] - Q[S]).$$

Equivalently, $d_{TV}(P,Q) = \frac{1}{2} \sum_{i=1}^{M} |P(i) - Q(i)| = \frac{1}{2}||P-Q||_1$

Definition 2. *The Kullback-Leibler (KL) divergence [2] (relative entropy) between P and Q is*

$$D(P||Q) = \sum_{i \in [M]} P(i) \log \left(\frac{P(i)}{Q(i)} \right).$$

Note that, $D(P||Q) \geq 0$ with equality if and only if $P = Q$.

Definition 3. *Let P and Q be joint distributions for random variables X and Y with support over $[M] \times [M]$. The Conditional KL Divergence [2] between P and Q is*

$$\begin{aligned} D(P_{Y|X}||Q_{Y|X}) &= \sum_{x \in [M]} P(x) D\left(P_{Y|X}(Y|X=x) || Q_{Y|X}(Y|X=x) \right) \\ &= \sum_x P(x) \sum_y P_{Y|X}(y|x) \log \left(\frac{P_{Y|X}(y|x)}{Q_{Y|X}(y|x)} \right) \end{aligned}$$

We have the following properties which relate these distances between distributions:

Property 1 (Pinsker's Inequality [3]). *For probability distributions P and Q on $[M]$,*

$$D(P||Q) \geq \frac{1}{2}||P-Q||_1^2$$

Property 2 (Chain rule of KL Divergence [3]). *Let P and Q be joint distributions for random variables X and Y with support over $[M] \times [M]$. Then,*

$$D(P_{XY}||Q_{XY}) = D(P_X||Q_X) + D(P_{Y|X}||Q_{Y|X})$$

References

[1] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Machine Learning*, 5(1):1–122, 2012.

[2] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[3] Imre Csiszar and János Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.