E1 245: Online Prediction & Learning	**Fall 2014**

Lecture 14 — September 18

Lecturer: Aditya Gopalan | *Scribe: Indu John*

14.1 Recap

In the last two classes, we studied the $EXP-3$ algorithm that enjoys a regret bound of $O\left(\sqrt{TNlogN}\right)$. Today, we establish a lower bound on regret for any algorithm in the adversarial bandit framework, which will imply that the upper bound cannot be improved beyond logarithmic factors.

14.2 Lower bound on regret in adversarial bandit framework

We will continue to deal with rewards(or gains) instead of losses, for the sake of convenience. The following theorem gives a lower bound on the regret of any prediction strategy(randomized or deterministic) in the adversarial(non-stochastic) multi armed bandit setting.

Theorem 14.1 (Minimax lower bound). *Let sup be the supremum over all distribution of rewards such that, for $i = 1,...,N$, the rewards $g(i,1), g(i,2),...,g(i,T) \in \{0,1\}$ are i.i.d., and let inf be the infimum over all algorithms playing $I_1, I_2, ..., I_T$. Then,*

$$inf \quad sup \quad \max_{i=1}^{N} \left(\mathbb{E}\left[\sum_{t=1}^{T} g(i,t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} g(I_t,t)\right]\right) \geq \frac{1}{20}\sqrt{TN}$$

where expectations are with respect to both the random generation of rewards and the internal randomization of the algorithm.

The general idea of the proof is as follows. After T time steps, at least one arm is pulled less than or equal to $\frac{T}{N}$ times. For this arm, one cannot differentiate between a Bernoulli of parameter $\frac{1}{2}$ and a Bernoulli of parameter $\frac{1}{2} + \sqrt{\frac{N}{T}}$. Thus if all arms are Bernoulli of parameter $\frac{1}{2}$ but one arm has parameter $\frac{1}{2} + \sqrt{\frac{N}{T}}$, then the algorithm should incur a regret of order $T\sqrt{\frac{N}{T}} = \sqrt{NT}$. To formalize this idea, we use the Kullback-Leibler divergence, and in particular Pinsker's inequality to compare the behavior of a given algorithm on the null bandit (where all arms are Bernoulli of parameter $\frac{1}{2}$) and the same bandit where we raise the parameter of one arm by ε.

We shall prove a more general lemma, which leads to Theorem 14.1 by a simple optimization over ε.

Lemma 14.2. *Let $\varepsilon \in [0,1]$. For any $i \in \{1,2,...,N\}$, let $\mathbb{E}_i$ denote the expectation against the joint distribution of rewards where all arms are i.i.d. Bernoulli of parameter $\frac{1-\varepsilon}{2}$ except arm i, which is i.i.d. Bernoulli of parameter $\frac{1+\varepsilon}{2}$. Then, for any algorithm,*

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

This implies that,

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

since max is always greater than the mean.

Proof: We shall prove the lemma in 5 steps, as given below.
Step I : Empirical distribution of plays
We start by considering a deterministic algorithm.
Let $S_{i,T} = \sum_{t=1}^{T}\mathbb{1}\{I_t = i\}$, the number of times arm i was played in T rounds.
Let $q_T := (q_{1,T}, q_{2,T}, ..., q_{N,T})$ be the empirical distribution of plays over the arms defined by $q_{i,T} = \frac{S_{i,T}}{T}$.
Let $J_T \sim q_T$. Then, $J_T \in \{1,2,...,N\}$. Let $\mathbb{P}_i$ be the probability mass function of J_T when all arms are i.i.d. Bernoulli of parameter $\frac{1-\varepsilon}{2}$ except arm i, which is i.i.d. Bernoulli of parameter $\frac{1+\varepsilon}{2}$. Observe that $\mathbb{E}_i\left[\frac{S_{j,T}}{T}\right] = \mathbb{P}_i[J_T = j]$. Hence,

$$\begin{aligned}
\mathbb{E}_i\left[\sum_{t=1}^{T}(g(i,t)-g(I_t,t))\right] &= \mathbb{E}_i\left[\sum_{t=1}^{T}\sum_{j\neq i}\mathbb{1}\{I_t = j\}(g(i,t)-g(I_t,t))\right] \\
&= \varepsilon\mathbb{E}_i\left[\sum_{t=1}^{T}\sum_{j\neq i}\mathbb{1}\{I_t = j\}\right] \\
&= \varepsilon T\sum_{j\neq i}\mathbb{E}_i\left[\frac{S_{j,T}}{T}\right] \\
&= \varepsilon T\sum_{j\neq i}\mathbb{P}_i(J_T = j) \\
&= \varepsilon T\left[1-\mathbb{P}_i(J_T = i)\right]
\end{aligned}$$

which implies

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_i\left[\sum_{t=1}^{T}(g(i,t)-g(I_t,t))\right] = \varepsilon T\left[1-\sum_{i=1}^{N}\mathbb{P}_i(J_T = i)\right] \tag{14.1}$$

Step II : Pinsker's inequality
Let $\mathbb{P}_0$ be the probability mass function of J_T when all the arms have the reward model i.i.d

Bernoulli$\left(\frac{1-\varepsilon}{2}\right)$. We use the following inequality to bound the RHS of equation (14.1).
Let μ, ν be two probability distributions on $\{1,2,...,N\}$. Then the KL divergence of ν from μ, $D(\mu||\nu)$ satisfies,

$$\begin{aligned} \sqrt{\frac{1}{2}D(\mu||\nu)} &:= \sqrt{\frac{1}{2}\sum_{i=1}^{N}\mu_i log\frac{\mu_i}{\nu_i}} \\ &\geq (\nu_i - \mu_i) \quad \forall i \end{aligned}$$

Thus,

$$\begin{aligned} \sqrt{\frac{1}{2}D(\mathbb{P}_0||\mathbb{P}_i)} &\geq \mathbb{P}_i[J_T = i] - \mathbb{P}_0[J_T = i] \\ \Rightarrow \sum_{i=1}^{N}\sqrt{\frac{1}{2}D(\mathbb{P}_0||\mathbb{P}_i)} &\geq \sum_{i=1}^{N}\mathbb{P}_i[J_T = i] - 1 \end{aligned}$$

$$\Rightarrow \frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_i[J_T = i] \leq \frac{1}{N} + \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{2}D(\mathbb{P}_0||\mathbb{P}_i)} \tag{14.2}$$

Step III : Computation of $D(\mathbb{P}_0||\mathbb{P}_i)$
In this step, we use some tools from information theory to derive an expression for the bound.
Note that, since the algorithm is deterministic, the sequence of observed rewards
$g^T = (g(I_1,1), g(I_2,2), ..., g(I_T,T))$ uniquely determines the empirical distribution of plays q_T. Let $\mathbb{P}_0^T$ be the pmf of g^T under the reward model where each arm's reward $\sim$ Bernoulli$\left(\frac{1-\varepsilon}{2}\right)$; and $\mathbb{P}_i^T$ be the pmf of g^T under the reward model where arm i's reward $\sim$ Bernoulli$\left(\frac{1+\varepsilon}{2}\right)$ and the rewards of other arms $\sim$ Bernoulli$\left(\frac{1-\varepsilon}{2}\right)$.
From information theory[1], we have $D(\mathbb{P}_0||\mathbb{P}_i) \leq D(\mathbb{P}_0^T||\mathbb{P}_i^T)$.

[1]Using chain rule for KL divergence.

Now, we can use the chain rule for KL divergence[2] as follows.

$$\begin{aligned}
D(\mathbb{P}_0^T||\mathbb{P}_i^T) &= D(\mathbb{P}_0^1||\mathbb{P}_i^1) + \sum_{t=2}^{T}\sum_{g^{t-1}} \left[\mathbb{P}_0^{t-1}(g^{t-1}) \times D\left(\mathbb{P}_0^t(\cdot|g^{t-1})||\mathbb{P}_i^t(\cdot|g^{t-1})\right)\right] \\
&= \mathbb{1}\{I_1 = i\} D\left(\frac{1-\varepsilon}{2}\bigg|\bigg|\frac{1+\varepsilon}{2}\right) + \mathbb{1}\{I_1 \neq i\}\, D\left(\frac{1-\varepsilon}{2}\bigg|\bigg|\frac{1-\varepsilon}{2}\right) \\
&\quad + \sum_{t=2}^{T}\left\{\sum_{g^{t-1}:I_t=i} \mathbb{P}_0^{t-1}(g^{t-1})\, D\left(\frac{1-\varepsilon}{2}\bigg|\bigg|\frac{1+\varepsilon}{2}\right) + \sum_{g^{t-1}:I_t\neq i} \mathbb{P}_0^{t-1}(g^{t-1})\, D\left(\frac{1-\varepsilon}{2}\bigg|\bigg|\frac{1-\varepsilon}{2}\right)\right\} \\
&= D\left(\frac{1-\varepsilon}{2}\bigg|\bigg|\frac{1+\varepsilon}{2}\right)\sum_{t=1}^{T}\sum_{g^{t-1}:I_t=i} \mathbb{P}_0^{t-1}(g^{t-1}) \qquad \left(\text{Since } D\left(\tfrac{1-\varepsilon}{2}\big|\big|\tfrac{1-\varepsilon}{2}\right) = 0\right) \\
&= D\left(\frac{1-\varepsilon}{2}\bigg|\bigg|\frac{1+\varepsilon}{2}\right)\sum_{t=1}^{T} \mathbb{P}_0[I_t = i] \\
&= D\left(\frac{1-\varepsilon}{2}\bigg|\bigg|\frac{1+\varepsilon}{2}\right)\mathbb{E}_0\left[S_{i,T}\right] \qquad (14.3)
\end{aligned}$$

Step IV : Conclusion for deterministic algorithms
Thus, we get

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\sqrt{D(\mathbb{P}_0||\mathbb{P}_i)} &\le \frac{1}{N}\sum_{i=1}^{N}\sqrt{D(\mathbb{P}_0^T||\mathbb{P}_i^T)} \\
&\le \sqrt{\frac{1}{N}\sum_{i=1}^{N} D(\mathbb{P}_0^T||\mathbb{P}_i^T)} \qquad \text{(Cauchy - Schwartz inequality)} \\
&= \sqrt{\frac{1}{N}\sum_{i=1}^{N} D\left(\frac{1-\varepsilon}{2}\bigg|\bigg|\frac{1+\varepsilon}{2}\right)\mathbb{E}_0\left[\frac{S_{i,T}}{T}\right]T} \qquad \text{(From (14.3))} \\
&= \sqrt{D\left(\frac{1-\varepsilon}{2}\bigg|\bigg|\frac{1+\varepsilon}{2}\right)\frac{T}{N}\sum_{i=1}^{N}\frac{1}{N}} \qquad \left(\mathbb{E}_0\left[\tfrac{S_{i,T}}{T}\right] = \tfrac{1}{N}\forall i\right) \\
&= \sqrt{\varepsilon\log\left(\frac{1+\varepsilon}{1-\varepsilon}\right)\frac{T}{N}}
\end{aligned}$$

Substituting in equation (14.1),

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_i\left[\sum_{t=1}^{T}\left(g(i,t) - g(I_t,t)\right)\right] \ge \varepsilon T\left[1 - \frac{1}{N} - \frac{1}{\sqrt{2}}\sqrt{\frac{T}{N}\varepsilon\log\left(\frac{1+\varepsilon}{1-\varepsilon}\right)}\right] \qquad (14.4)$$

We deduce the result from this by optimizing ε. We have,

$$\log\left(\frac{1+\varepsilon}{1-\varepsilon}\right) \approx \varepsilon - (-\varepsilon) = 2\varepsilon$$

[2] $D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x)) = D(p(x)||q(x)) + \sum_x p(x)D(p(\cdot|x)||q(\cdot|x))$

Setting $\varepsilon = c\sqrt{\frac{N}{T}}$, the RHS of (14.4) becomes,

$$\begin{aligned} c\sqrt{NT}\left(\frac{1}{2} - \frac{1}{\sqrt{2}}\sqrt{\frac{T}{N}\varepsilon \cdot 2\varepsilon}\right) &= c\sqrt{NT}\left(\frac{1}{2} - \sqrt{c}\right) \\ &= \Omega(\sqrt{NT}) \end{aligned}$$

Step V : Extend result to randomized algorithms

The result for deterministic algorithms can easily be extended to randomized algorithms. Let $\mathbb{E}_r$ denote the expectation with respect to the algorithm's internal randomization. Then, we have,

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_i\left[\sum_{t=1}^{T}\mathbb{E}_r\left(g(i,t) - g(I_t,t)\right)\right] = \mathbb{E}_r\left[\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_i\left[\sum_{t=1}^{T}\left(g(i,t) - g(I_t,t)\right)\right]\right]$$

Applying the lower bound on $\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_i\left[\sum_{t=1}^{T}\left(g(i,t) - g(I_t,t)\right)\right]$, and noticing that averaging the lower bounds preserves the lower bound, we obtain the desired result. $\square$

References

[1] Sebastien Bubeck and Nicolo Cesa-Bianchi, *Regret analysis of stochastic and nonstochastic multi-armed bandit problems, Theorem 3.5*. Foundations and Trends in Machine Learning, Vol.5, No.1, 2012.

[2] P. Auer, N. Cesa-Bianchi, Y. Freund and R. E. Schapire, *The nonstochastic multiarmed bandit problem*. SIAM Journal on Computing, Vol 32, No. 1, 2002.

[3] Nicolo Cesa-Bianchi and Gabor Lugosi, *Prediction, Learning and Games, Chapter 6, Section 6.9*. Cambridge University Press, 2006.