18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet　　Lecture 20
Scribe: Vira Semenova　　Nov. 23, 2015

In this lecture, we talk about the adversarial bandits under limited feedback. Adversarial bandit is a setup in which the loss function $l(a, z) : \mathcal{A}x\mathcal{Z}$ is determinitic. Limited feedback means that the information available to the DM after the step t is $\mathcal{I}_t = \{l(a_1, z_1), ..., l(a_{t-1}, z_t)\}$, namely consits of the realised losses of the past steps only.

5. ADVERSARIAL BANDITS

Consider the problem of prediction with expert advice. Let the set of adversary moves be $\mathcal{Z}$ and the set of actions of a decision maker $\mathcal{A} = \{e_1, ..., e_K\}$. At time t, $a_t \in \mathcal{A}$ and $z_t \in \mathcal{Z}$ are simultaneously revealed.Denote the loss associated to the decision $a_t \in \mathcal{A}$ and his adversary playing z_t by $l(a_t, z_t)$. We compare the total loss after n steps to the minimum expert loss, namely:

$$\min_{1 \le j \le K} \sum_{t=1}^{n} l_t(e_j, z_t),$$

where e_j is the choice of expert $j \in \{1, 2, .., K\}$.

The cumulative regret is then defined as

$$R_n = \sum_{t=1}^{n} l_t(a_t, z_t) - \min_{1 \le j \le K} \sum_{t=1}^{n} l_t(e_j, z_t)$$

The feedback at step t can be either full or limited. The full feedback setup means that the vector $f = (l(e_1, z_t), ..., l(e_K, z_t))^\top$ of losses incurred at a pair of adversary's choice z_t and each bandit $e_j \in \{e_1, ..., e_K\}$ is observed after each step t. Hence, the information available to the DM after the step t is $\mathcal{I}_t = \cup_{t'=1}^{t} \{l(a_1, z_t'), ..., l(a_K, z_{t'})\}$. The limited feedback means that the time $-t$ feedback consists of the realised loss $l(a_t, z_t)$ only. Namely, the information available to the DM is $\mathcal{I}_t = \{l(a_1, z_1), ..., l(a_t, z_t)\}$. An example of the first setup is portfolio optimization problems, where the loss of all possible portfolios is observed at time t. An example of the second setup is a path planning problem and dynamic pricing, where the loss of the chosen decision only is observed. This lecture has limited feedback setup.

The two strategies, defined in the past lectures, were exponential weights, which yield the regret of order $R_n \le c\sqrt{n \log K}$ and Follow the Perturbed Leader. We would like to play exponential weights, defined as:

$$p_{t,j} = \frac{\exp(-\eta \sum_{s=1}^{t-1} l(e_j, z_s))}{\sum_{l=1}^{k} \exp(-\eta \sum_{s=1}^{t-1} l(e_j, z_s))}$$

This decision rule is not feasible, since the loss $l(e_j, z_t)$ are not part of the feedback if $e_j \ne a_t$. We will estimate it by

$$\hat{l}(e_j, z_t) = \frac{l(e_j, z_t)\mathbb{1}(a_t = e_j)}{P(a_t = e_j)}$$

Lemma: $\hat{l}(e_j, z_t)$ is an unbiased estimator of $l(e_j, z_t)$

Proof. $E_{a_t}\hat{l}(e_j, z_t) = \sum_{k=1}^{K} \frac{l(e_k,z_t)\mathbb{1}(e_k=e_t)}{P(a_t=e_j)} P(a_t = e_k) = l(e_j, z_t)$ □

Definition (Exp 3 Algorithm): Let $\eta > 0$ be fixed. Define the exponential weights as

$$p_{t+1,j}^{(j)} = \frac{\exp(-\eta \sum_{s=1}^{t-1} \hat{l}(e_j, z_s))}{\sum_{l=1}^{k} \exp(-\eta \sum_{s=1}^{t-1} \hat{l}(e_j, z_s))}$$

($Exp3$ stands for Exponential weights for Exploration and Exploitation.)

We will show that the regret of Exp3 is bounded by $\sqrt{2nK \log K}$. This bound is $\sqrt{K}$ times bigger than the bound on the regret under the full feedback. The $\sqrt{K}$ multiplier is the price of have smaller information set at the time t. The are methods that allow to get rid of $\log K$ term in this expression. On the other hand, it can be shown that $\sqrt{2nK}$ is the optimal regret.

Proof. Let $W_{t,j} = \exp(-\eta \sum_{s=1}^{t-1} \hat{l}(e_j, z_s))$, $W_t = \sum_{j=1}^{k} W_{t,j}$, and $p_t = \frac{\sum_{j=1}^{K} W_{t,j} e_j}{W_t}$.

$$\log(\frac{W_{t+1}}{W_t}) = \log(\frac{\sum_{j=1}^{K} \exp(-\eta \sum_{s=1}^{t-1} \hat{l}(e_j, z_s)) \exp(-\eta \hat{l}(e_j, z_t))}{\sum_{j=1}^{K} \exp(-\eta \sum_{s=1}^{t-1} \hat{l}(e_j, z_s))}) \quad (5.1)$$

$$= \log(\mathbb{E}_{\mathcal{J}\sim p_t} \exp(-\eta \sum_{s=1}^{t-1} \hat{l}(e_{\mathcal{J}}, z_s))) \quad (5.2)$$

$$\leq^* \log(1 - \eta \mathbb{E}_{\mathcal{J}\sim p_t} \hat{l}(e_{\mathcal{J}}, z_s) + \frac{\eta^2}{2} \mathbb{E}_{\mathcal{J}\sim p_t} \hat{l}^2(e_{\mathcal{J}}, z_s) \quad (5.3)$$

where $*$ inequality is obtained by plugging in $\mathbb{E}_{\mathcal{J}\sim p_t}\hat{l}(e_{\mathcal{J}}, z_t)$ into the inequality

$$\exp x \geq 1 - \eta x + \frac{\eta^2 x^2}{2}$$

.

$$\mathbb{E}_{\mathcal{J}\sim p_t}\hat{l}(e_{\mathcal{J}}, z_t) = \sum_{j=1}^{K} p_{t,j} \hat{l}(e_{\mathcal{J}}, z_t) = \sum_{j=1}^{K} p_{t,j} \frac{l(e_j, z_t)\mathbb{1}(a_t = e_j)}{P(a_t = e_j)} = l(a_t, z_t) \quad (5.4)$$

$$\mathbb{E}_{\mathcal{J}\sim p_t}\hat{l}^2(e_{\mathcal{J}}, z_t) = \sum_{j=1}^{K} p_{t,j} \hat{l}^2(e_{\mathcal{J}}, z_t) = \sum_{j=1}^{K} p_{t,j} \frac{l^2(e_j, z_t)\mathbb{1}(a_t = e_j)}{P^2(a_t = e_j)} \quad (5.5)$$

$$= \frac{l^2(e_j, z_t)}{P_{a_t,t}} \leq \frac{1}{P_{a_t,t}} \quad (5.6)$$

Summing from 1 through n, we get
$\log(W_{t+1}) \leq \log(W_1) - \eta \sum_{t=1}^{n} l(a_t, z_t) + \frac{\eta^2}{2} \sum \frac{1}{P_{a_t,t}}$.

For $t = 1$, we initialize $w_{1,j} = 1$, so $W_1 = K$.

Since $\mathbb{E}_{\mathcal{J}} \frac{1}{P_{a_t,t}} = \sum_{j=1}^{K} \frac{p_{j,t}}{p_{j,t}} = K$, the expression above becomes

$\mathbb{E} \log(W_{n+1}) - \log K \leq -\eta \sum_{t=1}^{n} l(a_t, z_t) + \frac{\eta^2 K n}{2}$

Noting that $\log(W_{n+1}) = \log(\sum_{j=1}^{K} \exp(-\eta \sum_{s=1}^{t-1} \hat{l}(e_j, z_s))$

and defining $j^* = \mathrm{argmin}_{1 \leq j \leq K} \sum_{t=1}^{n} l(e_j, z_t)$, we obtain:

$$\log(W_{n+1}) \geq \log(\sum_{j=1}^{K} \exp(-\eta \sum_{s=1}^{t-1} l(e_j, z_s))) = -\eta \sum_{s=1}^{t-1} l(e_j, z_s)$$

Together:

$$\sum_{t=1}^{n} l(a_t, z_t) - \min_{1 \leq j \leq K} \sum_{t=1}^{n} l(e_j, z_t) \leq \frac{\log K}{\eta} + \frac{\eta K n}{2}$$

The choice of $\eta := \sqrt{2 \log K n K}$ yields the bound $R_n \leq \sqrt{2K \log K n}$. □