Machine Learning **Fall Semester, 2018/19**

Lecture 12: December 31, 2018

Lecturer: Yishay Mansour *Scribe: Inbar Oren, Michael Bargury*

1 Contexual Bandits

Stochastic Model:

- The environment sample, $x_t \sim D$ i.i.d. where $x_t \in X$.
- The learner, given x_t, chooses an action $a_t \in A$.
- The environment reveals $\ell_t \sim D(\ell|a_t, x_t)$.

Adverserial Model:

- The environment selects (adverserially) x_t and losses $\ell_t \in R^k$, where $x_t \in X$ and $k = |A|$.
- The learner, given x_t, chooses an action $a_t \in A$.
- The environment reveals $\ell_t(a_t)$ the loss of action a_t, i.e.

In each turn, the learner receives a context x_t, and chooses an action a_t. This implies that there will be a class of functions Π that receive x_t and returns a_t.

Let Π be a class of policies s.t

$$\forall \pi \in \Pi, \pi : X \to A.$$

The regret is defined as follows:

- Stochastic case:

$$\sum_{t=1}^{T} E[l_t(a_t)] - \min_{\pi \in \Pi} \sum_{t=1}^{T} E[l_t(\pi(x_t))]$$

  where the expectation is both over x_t, l_t and a_t.

- Adversarial case:

$$\sum_{t=1}^{T} E[l_t(a_t)] - \min_{\pi \in \Pi} \sum_{t=1}^{T} l_t(\pi(x_t))$$

  where the expectation is over a_t, and x_t, l_t are arbitrary.

For the usual MultiArm Bandit definition, define π_a to return the constant function that for any context returns action a, i.e.

$$\forall a \in A.\ \pi_a(x) = a$$

Based on scribe notes of Maya Schachter, Itai Admi and Nathanel Ozeri from 2018/17.

2 Small Number of Context

The small number of context case, is the "easy to handle case", but not the common one. In this case, $|X|$ is small and for each context $x \in X$ we will use a different MAB.

Algorithm: For each $x \in X$ we will use ALG_x with regret R_x.
In time t:

- Given x_t, we will run ALG_{x_t}.
- ALG_{x_t} returns an action a_t, we preform a_t and get loss ℓ_t.
- We return ALG_x the loss $\ell_t(a_t)$.

At time T let T_x be the number of times we had $x_t = x$. Note that,

$$\sum_x T_x = T.$$

Now, our regret R can be at most the sum of all regrets (recall that $k = |A|$):

$$R \leq \sum_{x \in X} R_x(T_x, k) = \sum_{x \in X} O(\sqrt{kT_x \log T}) \leq O(\sqrt{|X|kT \log T})$$

Notice we used the claim

$$\sum_{x \in X} \sqrt{T_x} \leq \sqrt{|X|T},$$

since the function is concave, the maximum value obtained when all the T_x are equal.

3 EXP4

- We have a set A of k actions.
- We have a set M of m experts.
- In each t: each expert i returns $q_{t,i} \in \Delta(A)$.

We can achieve a regret of $O(\sqrt{Tm \log T})$ by having the actions be the experts. Our goal is to obtain a regret of $O(\sqrt{Tk \log m})$.
The regret (compared to the best expert):

$$R = E[\sum_{t=1}^{T} l_t(a_t)] - \min_{i \in M} \sum_{t=1}^{T} l_t^T q_{t,i}$$

Let $S = min\{m, k\}$, then we will show a regret bound of,

$$Regret = O(\sqrt{TS \log m}).$$

The idea of EXP4, is to run EXP3, but in every stage when choosing an action a_t and receive a payoff p_t, we update all the experts that recommended on action a_t. We assume that at time t, each expert $i \in M$ gives a distribution $q_{t,i} \in \Delta(A)$.

Algorithm EXP4:

- Init $\forall i\ 1 \leq i \leq m.\ w_1(i) = 1$
- Time t:
  - $P_t(i) = \frac{w_t(i)}{W_t}$; $W_t = \sum_{i \in M} w_t(i)$

- Expert i gives $q_{t,i} \in \Delta(A)$
- $\forall a.\ Q_t(a) = \sum_{i \in M} P_t(i) q_{t,i}(a)$
- Draw action $a_t \sim Q_t$
- Observe loss $l_t(a_t)$
- Define

$$\hat{l}_t(a) = \begin{cases} \frac{l_t(a)}{Q_t(a)} & \text{if } a = a_t \\ 0 & \text{otherwise} \end{cases}$$

- Update $\forall i \in M.\ w_{t+1}(i) = w_t(i) e^{-\eta \hat{l}_t^T q_{t,i}}$. We define $h_t(i) = \hat{l}_t^T q_{t,i}$

3.1 Analysis

We show that $h_t(i)$ is an unbiased estimate of the loss of expert i. i.e. $loss_{t,i} = l_t^T q_{t,i}$.

Expert i can see the loss $\hat{l}_t^T q_{t,i} = h_t(i) = loss_{t,i}$,

$$E[h_t(i)|Q_t] = E[l_t^T q_{t,i}|Q_t] = \sum_a Q_t(a) \frac{q_{t,i}(a)\ l_t(a)}{Q_t(a)} = \sum_a q_{t,i}(a) l_t(a) = l_t^T q_{t,i} = loss_{t,i}.$$

If we will recall the regret proof for expert weights

$$E[\sum_{t=1}^T P_t^T h_t] - \sum_{t=1}^T h_t(i) \leq \frac{1}{\eta} \log m + \frac{\eta}{2} \sum_{t=1}^T \sum_{i=1}^m E[P_t(i) h_t^2(i)]$$

Now, we bound the second moment term,

$$\sum_{i=1}^m E[P_t(i) h_t^2(i)] \tag{1}$$

We start with

$$E[h_t^2(i)|P_t] = \sum_{a \in A} Q_t(a) (\frac{l_t(a) q_{t,i}(a)}{Q_t(a)})^2 \leq \sum_{a \in A} \frac{q_{t,i}^2(a)}{Q_t(a)}$$

Using the above, we can return to compute the bound:

$$\sum_{i=1}^m E[P_t(i) h_t^2(i)] = \sum_{i=1}^m E[P_t(i) E[h_t^2(i)|P_t]] \leq E[\sum_{a \in A} \frac{\sum_{i=1}^m P_t(i) q_{t,i}^2(a)}{Q_t(a)}]$$

Let

$$S_t(a) = \max_{i \in M} q_{t,i}(a)\ ;\ S_t = \sum_{a \in A} S_t(a)\ ;\ S = \max_t S_t$$

We get that

$$\sum_{i=1}^m E[P_t(i) h_t^2(i)] \leq E[\sum_{a \in A} \frac{\sum_{i=1}^m P_t(i) q_{t,i}(a)}{Q_t(a)} S_t(a)] = \sum_{a \in A} S_t(a) = S_t \leq S$$

Lets return to the general regret bound

$$E[Regret] = E[\sum_{t=1}^T P_t^T h_t] - \min_i \sum_{t=1}^T h_t(i) \leq \frac{1}{\eta} \log m + \frac{\eta}{2} TS$$

For $\eta = \sqrt{\frac{logm}{TS}}$ we have,

$$E[Regret] \leq O(\sqrt{TS \log m}).$$

Lets look at S:

- If all experts always agree, then $S_t = \sum_a q_{t,i}(a) = 1$ and we will get $O(\sqrt{T \log m})$

- $S_t(a) \leq 1 \implies S_t \leq k \implies S \leq k$ and we get a good Regret even if m is very large, but the number of the actions k is small. Namely, for $k \leq m$ we have regret $O(\sqrt{TK \log m})$.
- If the number of experts m is small, and the number of actions k is large we get

$$S_t(a) = \max_i q_{t,i}(a) \leq \sum_{i=1}^{m} q_{t,i}(a) S_t = \sum_{a \in A} S_t(a) \leq \sum_{i=1}^{m} \sum_{a \in A} q_{t,i}(a) = m$$

  And we get a regret of

$$O(\sqrt{mT \log m})$$

Although the regret of EXP4 is optimal, the main concern is that time and space complexity grows exponentially in the m, the number of experts.

4 Stochastic Contextual Bandits

Recall we defined Π to be the class of policies where $\pi \in \Pi$; $\pi : X \to A$.

We will first show a Greedy algorithm in the full information. For that, we will slightly change the model. Lets assume that after we choose an action, we can see the loss of all the actions.

Full Information Algorithm:

- At time t:
  - We calculate

$$\pi_t = \arg\min_{\pi \in \Pi} \sum_{\tau=1}^{t-1} l_\tau(\pi(x_t))$$

  - We use the action $a_t = \pi_t(x_t)$
  - We observe all loss functions $l_t(\cdot)$

Analysis: By using the Hoeffding bound, we give UCB and LCB bounds.
For i.i.d. random variables $X_t \in [0, 1]$, with probability $1 - \delta$:

$$|\frac{1}{t} \sum_{\tau=1}^{t} x_\tau - E[x_\tau]| \leq \sqrt{\frac{1}{t} \log \frac{2}{\delta}}$$

Theorem 1 *The policy π_t that is chosen in time t, satisfies w.p $1 - \delta$*

$$E[l_t(\pi_t(x_t))] \leq \min_{\pi \in \Pi} E[l_t(\pi(x_t))] + 2\sqrt{\frac{2}{t} \log \frac{2|\Pi|T}{\delta}},$$

where the expectation is both over the context x_t and the history (that determined π_t).

Proof: Let $\pi_t \in \Pi$ be the policy at time t. The policy π_t is a random variable, since is dependent on the history.
This implies we cannot argue directly about π_t. The idea would be to derive generalization bound for any time t and any policy. For every time t and any policy $\pi \in \Pi$ we have:

$$\left| \frac{1}{t-1} \sum_{\tau=1}^{t-1} l_\tau(\pi(x_\tau)) - E[l_\tau(\pi(x_\tau))] \right| \leq \sqrt{\frac{1}{2t} \log \frac{2|\Pi|T}{\delta}}$$

This implies that with probability $1 - \delta$, for every t:

$$\left| E[l_t(\pi_t(x_t))] - E[l_t(\pi^*(x_t))] \right| \leq 2\sqrt{\frac{1}{2t} \log \frac{2|\Pi|T}{\delta}}$$

□

Now, we compute the regret

$$E[Regret] \leq \sum_{t=1}^{T} \sqrt{\frac{2}{t} \log \frac{2|\Pi|T}{\delta}} = \sqrt{2 \log \frac{2|\Pi|T}{\delta}} \sum_{t=1}^{T} \frac{1}{\sqrt{t}} = O(\sqrt{T \log \frac{2|\Pi|T}{\delta}})$$

Since

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq \int_{1}^{T+1} \frac{1}{\sqrt{(x)}} dx = \left[2\sqrt{(x)}\right]_{1}^{T+1} = O(\sqrt{T})$$

5 Explore Exploit

- For $1 \leq t \leq B$:
  - Given a context x_t, ignore it and select uniformly at random action $a_t \in A$.
  - Preform a_t and get loss $l_t(a_t)$.
  - Keep the triplet $(x_t, a_t, l_t(a_t))$.
- Define π_B:

$$\pi_B = \arg\min_{\pi} \frac{1}{B} \sum_{t=1}^{B} l_t(a_t) \mathbb{1}(\pi(x_t) = a_t)$$

- For $t > B$:
  - Given x_t, select action $a_t = \pi_B(x_t)$ and receive loss $l_t(a_t)$

For the analysis we have,

$$\forall \pi \; E[l_t(a_t)\mathbb{1}(\pi(x_t) = a_t)] = E_{x_t, a_t}[E[l_t(a_t)\mathbb{1}(\pi(x_t) = a_t)|x_t]]$$

$$= \tfrac{1}{k} E_{x_t}[E[l_t(\pi(x_t))|x_t]] = \tfrac{1}{k} E(L(\pi))$$

Therefore,

$$\hat{R_B}(\pi) = E[\frac{1}{B} \sum_{t=1}^{B} l_t(a_t)\mathbb{1}(\pi(x_t) = a_t)] = \frac{1}{k} E(L(\pi))$$

We have an unbiased estimation for $E(L(\pi))$. We will perform Hoeffding bound. This implies that w.p. $1 - \delta$:

$$\left| \hat{R_B}(\pi) - \frac{E(L(\pi))}{k} \right| \leq \sqrt{\frac{1}{2B} \log \frac{2|\Pi|}{\delta}}$$

which implies,

$$L(\pi_B) \geq L(\pi^*) + 2k\sqrt{\frac{1}{2B} \log \frac{2|\Pi|}{\delta}}.$$

The bound for the regret is,

$$B + (T - B)(E(L(\pi_B)) - E(L(\pi^*))) \leq B + 2Tk\sqrt{\frac{2}{B} \log \frac{2|\Pi|}{\delta}} + \delta T.$$

For $\delta = \frac{1}{T}$ and

$$B = (Tk)^{\frac{2}{3}} (\log \frac{|\Pi|}{\delta})^{\frac{1}{3}},$$

we have,

$$E[Regret] = O((Tk)^{\frac{2}{3}} (\log(T|\Pi|))^{\frac{1}{3}}).$$

6 Lipschitz Bandits

We will assume a continuous action space $X = [0, 1]$ and loss which is L-Lipschitz

$|E[l(x)] - E[l(y)]| = |\mu(x) - \mu(y)| \leq L \cdot |x - y|$ for any two actions $x, y \in X$, where $\mu(x) = E[l(x)]$.

6.1 Simple Solution: discretization

We choose a finite set of arms $S \subset X$, $|S| = K$, of a grid of equal distance. Namely, for $\epsilon = \frac{1}{K+1}$, we have K points, $i\epsilon$, were $i \in \{1, 2, .., K\}$.

Let the lowest loss be,

$$\mu^*(X) = \min_{x \in X}(\mu(x))$$

The lowest loss over the set S be,

$$\mu^*(S) = \min_{x \in S}(\mu(x)).$$

We have a discretization error:

$$DE(S) = \mu^*(S) - \mu^*(X) \leq L \cdot \epsilon.$$

Using an algorithm ALG for stochastic MAB problem over the set S, we will get the regret bound:

$$E[regret(T)] = T \cdot (\mu_{ALG}(T) - \mu^*(X)) = \underbrace{T \cdot (\mu_{ALG}(T) - \mu^*(S))}_{regret_{ALG}(T,S)} + \underbrace{T(\cdot\mu^*(S) - \cdot\mu^*(X))}_{T \cdot DE(S)}$$

Assume that ALG has a regret:

$$regret_{ALG}(T, S) = O(\sqrt{TK \log K})$$

and we get:

$$E[regret(T)] \leq O(\sqrt{TK \log K}) + T \cdot L \cdot \epsilon = O(\sqrt{T \frac{1}{\epsilon} \log \frac{1}{\epsilon}}) + T \cdot L \cdot \epsilon$$

For $\epsilon = \left(\frac{\log TL}{L^2 T}\right)^{\frac{1}{3}}$ we get:

$$E[regret(T)] \leq O(T^{\frac{2}{3}} L^{\frac{1}{3}} \log^{\frac{1}{3}} TL)$$

6.2 Lower bound

We will show that $T^{\frac{2}{3}}$ is a lower bound on the regret by constructing a reduction from the MAB problem.

Recall that for profiles

$$I_j(i) = \begin{cases} Br(\frac{1}{2}) & i \neq j \\ Br(\frac{1}{2} - \epsilon) & i = j \end{cases}$$

$$I_0(i) = \frac{1}{2}$$

we have shown a lower bound of $\Omega(\sqrt{TK})$ for $\epsilon = O(\sqrt{(\frac{K}{T})})$.
We would like to translate this result to the Lipschitz Bandits model.

We will construct a set of $k = \frac{1}{\epsilon}$ profiles where the loss of the j'th profile is $\frac{1}{2}$ for all but a small interval, where it linearly changes to $\frac{1}{2} - \epsilon$ and back to $\frac{1}{2}$(see figure 1).

For every profile j, let $x_j = (2j - 1)\epsilon$. Define the profile's loss $\mu(x|I_j)$ as follows:

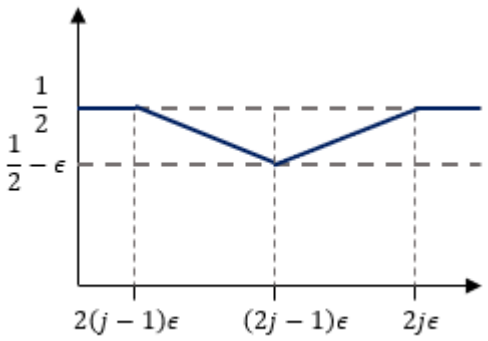

Figure 1: Loss of profile I_j

$$\mu(x|I_j) = \begin{cases} \frac{1}{2} & |x - x_j| \geq \epsilon \\ \frac{1}{2} - \epsilon + |x - x_j| & otherwise \end{cases}$$

Note that these losses are 1-Lipschitz.

Define S to be the set of point x_j. I.e. $S = \{x_j = (2j-1)\epsilon | j \in \{1, \ldots, k\}\}$.
For $x_i, x_j \in S$ we can rewrite $\mu(x|I_j)$ as

$$\mu(x_i|I_j) = \begin{cases} \frac{1}{2} & x_i, x_j \in S, x_i \neq x_j \\ \frac{1}{2} - \epsilon & x_i, x_j \in S, x_i = x_j \end{cases}$$

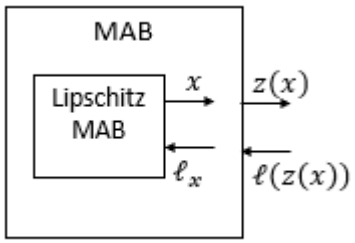

Figure 2: Solve MAB using Lipschitz MAB

Next, we use the MAB-Lipschitz algorithm to solve the MAB problem, which implies our lower bound. The details of the reduction are as follows (also see Figure 2).

1. First, for each $x \in R$ there exists an i such that $x \in [2(i-1)\epsilon, 2i\epsilon]$. Define a map $x \in R \mapsto z(x) \in S$ such that x is mapped to $z(x) = (2i-1)\epsilon \in S$. Notice that $|x - z(x)| \leq \epsilon$.

2. Given loss $l(z(x))$ from the discrete MAB problem, define $P_x = 1 - \frac{|x-z(x)|}{\epsilon}$ and

$$l_x = \begin{cases} l(z(x)) & \text{w.p } P_x \\ Br(\frac{1}{2}) & otherwise \end{cases}$$

Notice that we didn't use any knowledge of I_j, only $x, z(x)$ and $l(z(x))$.

Claim $E[l_x] = \mu(x)$
For intuition, think of the following cases:
- If $z(x) \neq x_j$ then $E[l(z(x))] = \frac{1}{2}$ and $l_x = Br(\frac{1}{2})$. Hence, $E[l_x] = \frac{1}{2}$.
- If $z(x) = x_j$ then $E[l(z(x))] = \frac{1}{2} - \epsilon$ and P_x changes l_x's expectation to the value we want it to be.

Proof:

$$E[l_x|x] = P_x \cdot \mu(z(x)|I_j) + (1 - P_x) \cdot \tfrac{1}{2} = \tfrac{1}{2} + P_x \cdot (\mu(z(x)|I_j) - \tfrac{1}{2}) = \begin{cases} \frac{1}{2} & z(x) \neq x_j \\ \frac{1}{2} + \epsilon P_x & z(x) = x_j \end{cases} = \mu(x)$$

□

At each time t and for each $j \in \{1, \ldots, k\}$ we get

$$\underbrace{E[\mu(a_t|I_j)]}_{\text{discrete MAB}} \leq \underbrace{E[\mu(x_t)]}_{\text{Lipschitz MAB}}$$

The loss is decreasing, and therefore regret is decreasing as well

$$E[Regret_{Lipschitz}(T)] \geq E[Regret_{MAB}(T)] \geq \sqrt{TK}$$

Where the last inequality is our lower bound on the MAB problem.
Furthermore, we have that $k = \frac{1}{\epsilon}$. Choosing $\epsilon = \frac{1}{2}T^{-\frac{1}{3}}$ yields $\sqrt{TK} \geq \Omega(T^{\frac{2}{3}})$. Putting it all together:

$$E[Regret_{Lipschitz}(T)] \geq \sqrt{TK} \geq \Omega(T^{\frac{2}{3}})$$

Which concludes our proof of the lower bound.