Machine Learning **Fall Semester, 2017/18**

Lecture 12: January 8, 2018

Lecturer: Yishay Mansour *Scribe: Maya Schachter, Itai Admi, Nathanel Ozeri*

1 Contexual Bandits

Stochastic Model:

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

Adverserial Model:

- The environment selects (adverserially) x_t and profits $r_t \in R^k$, where $x_t \in X$
- The learner, given x_t, chooses an action $a_t \in A$
- The environment reveals $r_t(a_t)$

In each turn, the lernear 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 regert is defined as follows:

- Stochastic case:

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

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

- Adversial case:

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

  where the expectation is over a_t.

For the usual Multiarm Bandit definition, define π_a to return the constant function for an action a, i.e.

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

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 we can fit it in the regret. For each Context $x \in X$ we will use MAB.

Algoritm: 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, and we preform it and get r_t
- We return ALG_x the gain $r_t(a_t)$

In 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:

$$R \leq \sum_{x \in X} R_x(T_x, K) = \sum_{x \in X} O(\sqrt{KT_x \log T}) \leq O(\sqrt{|X|KT \log T})$$

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 $q_{t,i} \in \Delta(A)$. where $q \in \{e_1 \dots e_m\}$.

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):

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

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 distribion $q_{t,i} \in \Delta(A)$.

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 $a_t \sim Q_t$
- Observe $l_t(a_t)$
- 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 $\forall i.\ w_{t+1,i} = w_t(i) e^{-\eta \hat{l}_t^T q_{t,i}}$. We define $h_t(i) = \hat{l}_t^T q_{t,i}$

4 Unbias estimate per expert

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

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

If we will recall the regret proof for expert weights

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

Now, we the second moment term,

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

We start with

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

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

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

Let

$$S_t(a) = \max_i q_{t,i}(a)\ ;\ S_t = \sum_a S_t(a)\ ;\ S = \max_t S_t$$

We get that

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

Lets return to the general regret bound

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

Per $\eta = \sqrt{\frac{logM}{TS}}$ 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 big, but the number of the actions k is small.

- If the number of experts M is small, and the number of actions A is big we get

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

And we get

$$O(\sqrt{MT \log M})$$

Although the regret is great for EXP4, the main concern is that time and space complexity grows exponentially in the experts.

5 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
  $$\pi_t = \arg\min_{\pi} \sum_{\tau=1}^{t-1} l_\tau(\pi(x_t))$$
  - 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 $X_t \in [0,1]$ i.i.d. with probability $1-\delta$:

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

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

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

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

Proof: Let $\pi \in \Pi$ be a policy. The random variable $l(\pi_\tau)$ is dependent on the history. we will demand "convergence" $\forall t$. and $\forall \pi$:

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

$$\underset{w.p\ 1-\delta}{\forall t}.\ \left|E[l(\pi_\tau)] - E[l(\pi^*)]\right| \leq 2\sqrt{\tfrac{1}{2t}\log\tfrac{2|\Pi|T}{\delta}}$$

$\square$

Now, let us compute the regret

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

While we know

$$\sum_{t=1}^{T}\frac{1}{\sqrt{t}} = O(\sqrt{T})$$

6 Explore Exploit

- For $1 \leq t \leq B$:
  - Given x_t, ignore it and uniformly choose action $a_t \in A$.
  - Preform a_t and get $r_t = l_t(a_t)$.
  - Keep (x_t, a_t, r_t).
- Define π_B:

$$\pi_B = \arg\min_{\pi} \sum_{x_t, a_t, r_t} r_t \mathbb{1}(\pi(x_t) = a_t)$$

- For $t > B$:
  - Given x_t, compute $\pi_B(x_t)$ and receive r_t

From here we get

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

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

Therefore,

$$\hat{R_B}(\pi) = \sum_{\tau=1}^{B} r_\tau \mathbb{1}(\pi(x_t) = a_t) = \frac{B}{k} E(L(\pi))$$

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

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

which implies,

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

The bound over the Regret is,

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

For

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

we have,

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

7 Lipschitz Bandits

We will assume a continuos arms space $X = [0, 1]$ which is L-Lipschitz

$$|E[r(x)] - E[r(y)]| = |\mu(x) - \mu(y)| \leq L \cdot |x - y| \text{ for any two arms } x, y \in X$$

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 best reward be,

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

The best reward over the set S be,

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

We have a discretization error:

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

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

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

Assume that ALG has a regret:

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

and we get:

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

We can choose$\epsilon = \left(\frac{\log T}{L^2 T}\right)^{\frac{1}{3}}$ and get:

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

Lower bound

We will see that $T^{\frac{2}{3}}$ is optimal in the worst case:

In a problem instance I_j were $j \in \{1, 2, .., K\}$, and x_j is the arm with the highest mean reward, with $L = 1$ and $0 - 1$ rewards, such that $\mu(x_j) = \frac{1}{2} + \varepsilon$ and all arms have mean reward:

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

This problem instance satisfies 1-Lipschitz condition.

Given ε we can choose $K = \frac{1}{2\varepsilon} + 1$ and for $x_i, x_j \in S = \{0, 2\varepsilon, 4\varepsilon, ..., 2K\varepsilon = 1\}$, we can define:

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

We will show a simulation that allows using the Lipschitz algorithm to solve the MAB problem for I_j inputs:

1. Choose $x \in [\frac{a}{K} - \varepsilon, \frac{a}{K} + \varepsilon]$ by the Lipschitz algorithm
2. Choose arm a (closest to x) by the MAB algorithm
3. Recieve reward r for the MAB algorithm with $E(r) = \mu_I(a)$
4. Recieve reward r_x for the Lipschitz algorithm with $E(r_x) = \mu(x)$

We can define a probability for r_x :

$$P_x = 1 - \frac{|x - \frac{a}{K}|}{\epsilon}$$

$$r_x = \begin{cases} r & w.p.P_x \\ Br(\frac{1}{2}) & otherwise \end{cases}$$

$$E(r_x|x) = P_x \cdot \mu_I(a) + (1 - P_x) \cdot \frac{1}{2} = \frac{1}{2} + P_x \cdot (\mu_I(a) - \frac{1}{2}) = \begin{cases} \frac{1}{2} & a \neq a^* \\ \frac{1}{2} + \varepsilon P_x & a = a^* \end{cases} = \mu(x)$$

Because in each time t the reward of the Lipschitz algorithm is bounded by the MAB algorithm, we get:

$$E[\mu_I(a_t)|x_t] \geq E[\mu(x_t)]$$

This implies that the regret of the Lipschitz MAB is not smaller then the MAB on S:

$$E[Regret_{Lipschitz}(T)|I] \geq E[Regret_{MAB}(T)|I]$$

Now we can use the lower bound for the MAB on S:

$$E_I[E[Regret_{Lipschitz}(T)|I]] \geq \sqrt{KT}.$$

To complete the proof, for $\varepsilon = \frac{1}{2}T^{-\frac{1}{3}}$ we get $K = T^{\frac{1}{3}}$ and $E_I[E[Regret_{Lipschitz}(T)|I]] \geq T^{\frac{2}{3}} = \Omega(\sqrt{KT})$.