Reinforcement Learning **Spring Semester, 2018/9**

Lecture 10: May 22, 2019

Lecturer: Yishay Mansour *Scribe: ym*

[1]

10.1 Stochastic Bandits and Regret Minimization

We consider a simplified model of an MDP where there is only a single state and a fixed set A of k actions (a.k.a., arms). We consider a finite horizon problem, where the horizon is T. Clearly, the planning problem is trivial, simply select the action with the highest expected reward. We will concentrate on the learning perspective, where the expected reward of each action are unknown.

At each round $1 \leq t \leq T$ the player selects and executes an action. After executing the action, the player observes the reward of the action. However, the rewards of the other actions in A are not revealed to the player.

The reward for action i at round t is denoted by $r_t(i) \sim D_i$, where the support of the reward distribution D_i is $[0, 1]$. We assume that the rewards are i.i.d. (independent and identically distributed).

Motivation

1. **News**: a user visits a news site and is presented with a news header. The user either clicks on this header or not. The goal of the website is to maximize the number of clicks. So each possible header is an action in a bandit problem, and the clicks are the rewards

2. **Medical Trials**: Each patient in the trial is prescribed one treatment out of several possible treatments. Each treatment is an action, and the reward for each patient is the effectiveness of the prescribed treatment.

3. **Ad selection**: In website advertising, a user visits a webpage, and a learning algorithm selects one of many possible ads to display. If an advertisement is

[1] Based on scribes of lecture 7 of Advanced Topics in ML and AGT, Fall Semester, 2017/18 and 2018/19

displayed, the website observes whether the user clicks on the ad, in which case the advertiser pays some amount $v_a \in [0,1]$. So each advertisement is an action, and the paid amount is the reward.

Model

- A set of actions $A = \{a_1 \ldots, a_k\}$
- Each action a_i has a reward distribution D_i over $[0,1]$.
- The expectation of distribution D_i is:

$$\mu_i = E_{X \sim D_i}[X]$$

- $\mu^* = \max_i \mu_i$ and $a^* = \arg\max_i \mu_i$.
- a_t is the action the learner chose at round t

$$Regret = \max_{i \in A} \sum_{t=1}^{T} \underbrace{r_t(i)}_{\text{Random variable}} - \sum_{t=1}^{T} \underbrace{r_t(a_t)}_{\text{Random variable}}$$

$$\begin{aligned} \text{Pseudo Regret} &= \max_i E\left[\sum_{t=1}^{T} r_t(i)\right] - E\left[\sum_{t=1}^{T} r_t(a_t)\right] \\ &= \mu^* \cdot T - \sum_{t=1}^{T} \mu_{a_t} \end{aligned}$$

We will use extensively the following concentration bound.

Theorem 10.1 (Hoeffding's inequality)**.** *Given $X_1, \ldots, X_m$ i.i.d random variables s.t $X_i \in [0,1]$ and $E[X_i] = \mu$.*

$$Pr[\underbrace{\frac{1}{m}\sum_{i=1}^{m} X_i - \mu}_{\frac{1}{m}S} \geq \epsilon] \leq \exp(-\tfrac{\epsilon^2 m}{2})$$

10.1.1 Warmup: Full information $k = 2$

We start with a simple case where there are two actions and we observe the reward of both actions at each time t. We will analyze the greedy policy, which selects the action with the higher average reward (so far).

The greedy policy at time t does the following:

- We observe $\langle r_t(1), r_t(2) \rangle$
- Define
$$avg_t(i) = \frac{1}{t} \sum_{\tau=1}^{t} r_\tau(i)$$
- In time $t+1$ we choose:
$$a_{t+1} = \arg \max_{i \in \{1,2\}} avg_t(i)$$

We now would like to compute the expected regret of the greedy policy. W.l.o.g., we assume that $\mu_1 \geq \mu_2$, and define $\Delta = \mu_1 - \mu_2 \geq 0$.

$$\text{Pseudo Regret} = \sum_{t=1}^{\infty} (\mu_1 - \mu_2) \Pr\left[avg_t(2) \geq avg_t(1)\right]$$

Clearly, at any time t,

$$E[avg_t(2) - avg_t(1)] = \mu_2 - \mu_1 = -\Delta$$

We can define a random variable $X_t = r_t(2) - r_t(1) + \Delta$ and $E[X_t] = 0$. Since $(1/t) \sum_t X_t = avg_t(2) - avg_t(1) + \Delta$, by Theorem 10.1

$$Pr[S_2(t) \geq S_1(t)] = Pr\left[avg_t(2) - avg_t(1) + \Delta \geq \Delta\right] \leq e^{-\Delta^2 \frac{t}{2}}$$

We can now bound the regret as follows,

$$\begin{aligned} E\left[\text{Pseudo Regret}\right] &= \sum_{t=1}^{\infty} \Delta \Pr\left[S_2(t) \geq S_1(t)\right] \\ &\leq \sum_{t=1}^{\infty} \Delta e^{-\Delta^2 \frac{t}{2}} \\ &\leq \int_0^{\infty} \Delta e^{-\Delta^2 \frac{t}{2}} dt \\ &= \left[\frac{2}{\Delta} e^{-\Delta^2 \frac{t}{2}}\right]_0^{\infty} \\ &= \frac{2}{\Delta} \end{aligned}$$

Notice that this bound does not depend on T!

10.1.2 Stochastic Multi-Arm Bandits

We will now see that we cannot get a regret that does not depend on T for the bandits case. Considering the following example:

$$a_1 \sim Br\left(\frac{1}{2}\right)$$

For action a_2 there are two alternatives, each with probability $1/2$,

$$a_2 \sim Br\left(\frac{1}{4}\right)\left(w.p.\frac{1}{2}\right) \qquad or \qquad a_2 \sim Br\left(\frac{3}{4}\right)\left(w.p.\frac{1}{2}\right)$$

Assume by way of contradiction

$$E\left[\sum_{i\in\{1,2\}} \Delta_i |T_i|\right] = E\left[PseudoRegret\right] = R$$

where R does not depend on T.
By Markov inequality:

$$Pr\left[PseudoRegret \geq 2R\right] \leq \frac{1}{2}$$

Since μ_1 is known, an optimal algorithm will first check a_2 in order to decide which action is better and stick with it.

Assuming $\mu_2 = \frac{1}{4}$, and the algorithm decided to stop playing a_2 after M rounds, Then:

$$PseudoRegret = \frac{1}{4}M$$

Thus,

$$Pr\left[PseudoRegret \geq 2R\right] = Pr\left[M \geq 8R\right] \leq \frac{1}{2}$$

And,

$$Pr\left[M < 8R\right] > \frac{1}{2}$$

Hence, the probability that after $8R$ rounds, the algorithm will stop playing a_2 (if $\mu_2 = \frac{1}{4}$) is at least $\frac{1}{2}$. This implies that there is some sequence of $8R$ outcomes which will result is stopping to try action a_2. For simplicity, assume that the sequence is the all zero sequence.

Assume $\mu_2 = \frac{3}{4}$, but all $8R$ first rounds, playing a_2 yield the value zero (with probability $\left(\frac{1}{4}\right)^{8R}$). We assumed that after $8R$ zeros for action a_2 the algorithm will stop playing a_2, even though it is the preferred action. In this case, we will get:

$$PseudoRegret = \frac{1}{4}(T - M) \approx \frac{1}{4}T$$

The expected Pseudo Regret is,

$$E\left[PseudoRegret\right] = R \geq \underbrace{\frac{1}{2}}_{a_2 \sim Pr\left(Br\left(\frac{3}{4}\right)\right)} \cdot \underbrace{\left(\frac{1}{4}\right)^{8R}}_{Pr(all\ 0 | a_2 \sim Pr\left(Br\left(\frac{3}{4}\right)\right))} \cdot (T - 8R) \approx e^{-O(R)}T$$

Which implies that:

$$R = O\left(\log T\right)$$

Contrary to the assumption that R does not depend on T.

10.2 Explore-Then-Exploit

1. We choose a parameter M. For M phases we choose each action once (for a total of kM rounds of exploration).

2. After kMrounds we always choose the action that had highest average reward during the explore phase.

Define:

$$\begin{aligned} T_j &= \{t : a_t = j, t \leq k \cdot M\} \\ \hat{\mu}_j &= \frac{1}{M} \sum_{t \in T_j} r_j(t) \\ \mu_j &= E[r_j(t)] \\ \Delta_j &= \mu^* - \mu_j \end{aligned}$$

where Δ_j is the difference in expected reward of action j and the optimal action.

We can now write the regret as a function of those parameters:

$$E\left[\text{Pseudo regret}\right] = \underbrace{\sum_{j=1}^{k} \Delta_j \cdot M}_{Explore} + \underbrace{(T - k \cdot M) \sum_{j=1}^{k} \Delta_j Pr\left[j = \arg\max_i \hat{\mu}_i\right]}_{Exploit}$$

For the analysis define:

$$\lambda = \sqrt{\frac{8 \log T}{M}}$$

By Theorem 10.1 we have

$$\Pr\left[|\hat{\mu}_j - \mu_j| \geq \lambda\right] \leq 2e^{-\frac{\lambda M}{2}} = \frac{2}{T^4}$$

which implies (using the union bound) that

$$\Pr\underbrace{\left[\exists_j : |\hat{\mu}_j - \mu_j| \geq \lambda\right]}_{B} \leq \frac{2k}{T^4} \underset{for\, k \leq T}{\leq} \frac{2}{T^3}$$

Define the "bad event" $B = \{\exists_j : |\hat{\mu}_j - \mu_j| \geq \lambda\}$. If B did not happen then for each action j, such that $\hat{\mu}_j \geq \hat{\mu}^*$, we have

$$\mu_j + \lambda \geq \hat{\mu}_j \geq \hat{\mu}^* \geq \mu^* - \lambda$$

therefore:

$$2\lambda \geq \mu^* - \mu_j = \Delta_j$$

and therefore:

$$\Delta_j \leq 2\lambda$$

Then, we can bound the expected regret as follows:

$$E[Regret] \leq \underbrace{\left(\sum_{j=1}^{k} \Delta_j\right) M}_{Explore} + \underbrace{(T - k \cdot M) \cdot 2\lambda}_{\text{B didn't happen}} + \underbrace{\frac{2}{T^3} \cdot T}_{\text{B happened}}$$

$$\leq k \cdot M + 2 \cdot \sqrt{\frac{8 \log T}{M}} \cdot T + \frac{2}{T^2}$$

If we optimize the number of exploration phases M and choose $M = T^{\frac{2}{3}}$, we get:

$$k \cdot T^{\frac{2}{3}} + 2 \cdot \sqrt{8 \log T} \cdot T^{\frac{2}{3}} + \frac{2}{T^2}$$

which is sub-linear but more than the $O(\sqrt{T})$ rate we would expect.

10.3 Improved Regret Minimization Algorithms

We will look at some more advanced algorithms that mix the exploration and exploitation.

Define:

$n_t(i)$ - the number of times we chose action i by round t

$\hat{\mu}_t(i)$ - the average reward of action i so far, that is:

$$\hat{\mu}_t(i) = \sum_{t=1}^{T} r_i(t) \mathbb{I}\left(a_t = i\right) \frac{1}{n_i(t)}$$

Notice that $n_i(t)$ is a random variable and not a number!

We would like to get the following result:

$$\Pr\left[|\hat{\mu}_t(i) - \mu_i| \leq \underbrace{\sqrt{\frac{8 \log T}{n_i(t)}}}_{\lambda_t(i)}\right] \geq 1 - \frac{2}{T^4}$$

We would like to look at the m^{th} time we sampled action i:

$$\hat{\mathbb{V}}_m(i) = \frac{1}{m} \sum_{\tau=1}^{m} r_i(t_\tau)$$

Where the t_τ's are the rounds when we chose action i

Now we fix m and get:

$$\forall i \forall m \quad \Pr\left[\left|\hat{\mathbb{V}}_m(i) - \mu_i\right| \leq \sqrt{\frac{8\log T}{m}}\right] \geq 1 - \frac{2}{T^4}$$

and notice that $\hat{\mu}_t(i) \equiv \hat{\mathbb{V}}_i(m)$ when $m = n_i(t)$.

Define the "good event" G:

$$G = \{\forall_i \forall_t \, |\hat{\mu}_i(t) - \mu_i| \leq \lambda_i(t)\}$$

The probability of G is,

$$Pr\,(G) \geq 1 - \frac{2}{T^2}$$

10.4 Refine Confidence Bound

Define the upper confidence bound:

$$UCB_t(i) = \hat{\mu}_t(i) + \lambda_t(i)$$

and similarly, the lower confidence bound:

$$LCB_t(i) = \hat{\mu}_t(i) - \lambda_t(i)$$

if G happened then:

$$\forall i \forall t \quad \mu_i \in [LCB_t(i), UCB_t(i)]$$

Therefore:

$$Pr\left[\forall i \forall t \quad \mu_i \in [LCB_t(i), UCB_t(i)]\right] \geq 1 - \frac{2}{T^2}$$

10.4.1 Successive Elimination

We maintain a set of actions S.

Initially $S = A$

In each phase:

- We try every $i \in S$ once

- For each $j \in S$ if there exists $i \in S$ such that:

$$UCB_t(j) < LCB_t(i)$$

  We remove j from S, that is we update:

$$S \leftarrow S - \{j\}$$

We will get the following results:

- As long as action i is still in S, we have tried action i exactly the same number of times as all of any other action $j \in S$.
- The best action, under the assumption that the event G holds, is never eliminated from S.

Under the assumption of G we get:

$$\mu^* - 2\lambda \leq \hat{\mu}^* - \lambda = LCB_* < UCB_i = \hat{\mu}_i + \lambda \leq \mu_i + 2\lambda$$

Where $\lambda = \lambda_i = \lambda^*$ because we have chosen action i and the best action the same number of times so far.

Therefore, assuming event G holds,

$$\Delta_i = \mu^* - \mu_i \leq 4\lambda = 4\sqrt{\frac{8\log T}{n_t(i)}}$$

$$\Rightarrow \quad n_T(i) \leq \frac{c}{\Delta_i^2}\log T$$

This implies that

$$\begin{aligned} E\left[\text{Pseudo Regret}\right] &= \sum_{i=1}^{k} \Delta_i n_i(t) \\ &\leq \sum_{i=1}^{k} \frac{c}{\Delta_i}\log T + \underbrace{\frac{2}{T^2}\cdot T}_{\text{The bad event}} \end{aligned}$$

meaning that the expected pseudo regret is bounded by $O\left(\frac{1}{T}\right)$.

10.4.2 Upper confidence bound (UCB)

The UCB algorithm simply uses the UCB bound. The algorithm works as follows:

- We try each action once (for a total of k rounds)
- Afterwards we choose:

$$a_t = \arg\max_i UCB_t(i)$$

If we chose action i then, assuming G holds, we have

$$UCB_t(i) \geq UCB_t(a^*) \geq \mu^*$$

where a^* is the optimal action.

Using the definition of UCB and the assumption that G holds, we have

$$UCB_t(i) = \hat{\mu}_t(i) + \lambda_t(i) \leq \mu_i + 2\lambda_t(i)$$

Since we selected action i at time t we have

$$\mu_i + 2\lambda_t(i) \geq \mu^*$$

Rearranging, we have,

$$2\lambda_t(i) \geq \mu^* - \mu_i = \Delta_i$$

Each time we chosen action i, we could not have made a very big mistake because:

$$\Delta_i \leq \ 2 \cdot \sqrt{\frac{8 \log T}{n_t(i)}}$$

And therefore if i is very far off from the optimal action we would not choose it too many times. We can bound the number of time action i is used by,

$$n_t(i) \leq \frac{c}{\Delta_i^2} \log T$$

And over all we get:

$$\begin{aligned} E\left[\text{Pseudo Regret}\right] &= \sum_{i=1}^{k} \Delta_i E\left[n_t(i)\right] + \underbrace{\frac{2}{T^2} \cdot T}_{\text{The bad event}} \\ &\leq \sum_{i=1}^{k} \frac{c}{\Delta_i} \cdot \log T + \frac{2}{T} \end{aligned}$$

10.5 Best Arm Identification

We would like to identify the best action, or an almost best action. We can define the goal in one of two ways.

PAC criteria An action i is ϵ-optimal if V. The PAC ctriteria is that, given $\epsilon, \delta > 0$,, with probability at least $1-\delta$, find an ϵ optimal action.

Exact identification Given $\Delta \leq \mu^* - \mu_i$ (for every suboptimal action i), find the optimal action a_*, with probability at least $1-\delta$.

10.5.1 Naive Algorithm (PAC criteria):

We sample each action i for $m = \frac{8}{\epsilon^2}\log\frac{2k}{\delta}$ times, and return $a = \arg\max_i \hat{\mu}_i$.

For rewards in $[0,1]$, then, by Theorem 10.1, for every action i we have

$$Pr\left[\underbrace{|\hat{\mu}_i - \mu_i| > \frac{\epsilon}{2}}_{\text{bad event}}\right] \leq 2e^{-\left(\frac{\epsilon}{2}\right)^2 m/2} = \frac{\delta}{k}$$

By union bound we get:

$$Pr\left[\exists_i\, |\hat{\mu}_i - \mu_i| > \frac{\epsilon}{2}\right] \leq \delta$$

If the bad event $B = \{\exists_i\, |\hat{\mu}_i - \mu_i| > \frac{\epsilon}{2}\}$ did not happen, then both: (1) $\mu^* - \frac{\epsilon}{2} \leq \hat{\mu}^*$ and (2) $\mu_i + \frac{\epsilon}{2} \leq \hat{\mu}_i$.

This implies,

$$\Rightarrow\ \mu_i + \frac{\epsilon}{2} \geq \hat{\mu}_i \geq \hat{\mu}^* \geq \mu^* - \frac{\epsilon}{2}$$

$$\Rightarrow\ \epsilon \geq \mu^* - \mu_i$$

And therefore $a = \arg\max_i \hat{\mu}_i$ is the optimal action in probability $1-\delta$

We would like to slightly improve the sample size of this algorithm

10.5.2 Median Algorithm

The idea: the algorithm runs for l phases, after each phase we eliminate half of the actions. This elimination allows us to sample each action more times in the next phase which makes eliminating the optimal action less likely.

Input: $\epsilon, \delta > 0$
Output: $\bar{a} \in A$
Init: $S_1 = A$, $\epsilon_1 = \frac{\epsilon}{4}$, $\delta_1 = \frac{\delta}{2}$, $l = 1$
Repeat:
$\forall_i \in S_l$, sample action i, $\frac{1}{\left(\frac{\epsilon_l}{2}\right)^2} \log\left(\frac{3}{\delta_l}\right)$ times
$\hat{\mu}_i \leftarrow$ mean (only of samples during the l^{th}phase)
$\text{median}_l \leftarrow median\left\{\hat{\mu}_i : i \in S_l\right\}$
$S_{l+1} \leftarrow \left\{i \in S_l : \hat{\mu}_i \geq \text{median}_l\right\}$
$\epsilon_{l+1} \leftarrow \frac{3}{4}\epsilon_l$
$\delta_{l+1} \leftarrow \frac{\delta_l}{2}$
$l \leftarrow l + 1$
Until $|S_l| = 1$

Algorithm 1: Best Arm Identification

Complexity: During phase l we have $|S_l| = \frac{k}{2^{l-1}}$ actions.

$$\epsilon_l = \frac{3}{4}\epsilon_{l-1} = \frac{\epsilon}{4}\left(\frac{3}{4}\right)^{l-1}, \ \delta_l = \frac{\delta}{2^l}$$

$$\Rightarrow \ \sum \epsilon_l \leq \epsilon, \ \sum \delta_l \leq \delta$$

The total number of samples is therefore:

$$ccc \sum_l |S_l| \cdot \frac{4}{\epsilon_l^2} \log \frac{3}{\delta_l} = \sum_l \frac{k}{2^{l-1}} \frac{64}{\epsilon^2} \left(\frac{16}{9}\right)^{l-1} \log \frac{3 \cdot 2^l}{\delta}$$

$$= \sum_l k \left(\frac{8}{9}\right)^{l-1} \left[c \cdot \frac{\log \frac{1}{\delta}}{\epsilon^2} + \frac{\log 3}{\epsilon^2} + \frac{l}{\epsilon^2}\right]$$

$$= O\left(\frac{k}{\epsilon^2} \log \frac{1}{\delta}\right)$$

Correctness:

Theorem 10.2. $Pr\left[\underbrace{\max_{j \in S_l} \mu_j}_{action\ l} \leq \underbrace{\max_{j \in S_{l+2}} \mu_j}_{action\ l+1} + \epsilon_l\right] \geq 1 - \delta_l$

Proof. We do the proof for $l = 1$, general l is similar. Define $E_1 = \left\{\hat{\mu}^* < \mu^* - \frac{\epsilon_1}{2}\right\}$. We have $Pr\left[E_1\right] \leq \frac{\delta_1}{3}$. If E_1 did not happen, we define a bad set:

$$\text{Bad} = \{j \ : \ \mu^* - \mu_j \geq \epsilon_1, \ \hat{\mu}_j \geq \hat{\mu}^*\}$$

Consider an action j such that $\mu^* - \mu_j \geq \epsilon_1$, then:

$$Pr[\hat{\mu}_j \geq \hat{\mu}^* | \underbrace{\hat{\mu}^* \geq \mu^* - \frac{\epsilon_1}{2}}_{\neg E_1}] \leq Pr[\hat{\mu}_j \geq \mu_j + \frac{\epsilon_1}{2} | \neg E_1] \leq \frac{\delta_1}{3}$$

Note that the probability is not negligible. We will show that it cannot happen to too many such actions. We will bound the expectation of the size of Bad,

$$E[|\text{Bad}||\neg E_1] \leq k\frac{\delta_1}{3}$$

with Markov's inequality we get:

$$Pr\left[|\text{Bad}| \geq \frac{k}{2}\middle|\neg E_1\right] \leq \frac{E\left|\text{Bad}\right|}{k/2} = \frac{2}{3}\delta_1$$

with probability $1 - \delta_1$: $\hat{\mu}^* \geq \mu^* - \frac{\epsilon_1}{2}$ and $|\text{Bad}| \leq \frac{k}{2}$. Therefore: $\exists_j \notin \text{Bad}$ and $j \in S_{l+1}$. □