CS 475/675 Machine Learning **Spring 2023**

Lecture 15: Stochastic Multi-Armed Bandit Algorithms

Lecturer: Thanh Nguyen-Tang *Scribe: Thanh Nguyen-Tang*

Acknowledgement: *LaTeX template courtesy of UC Berkeley EECS dept.*

We consider stochastic multi-armed bandits with (finite) action sets $[k] := \{1, \dots, k\}$ for some finite integer k. Each action $a \in [k]$ is regarded as an arm that the player plays/pulls. The interaction protocol proceeds as follows: For every time step $t = 1, 2, \dots, n$,

- The player chooses an *action* $a_t \in [k]$;
- The *reward* vector $z_t \in \mathbb{R}^k$ whose each component is sampled independently from (possibly different) unknown distributions. Namely, $z_t = (z_{t,1}, \dots, z_{t,k})$ where $z_{t,a}$ is a random number sampled from some unknown reward distribution p_a with mean μ_a for any $a \in [k]$. For simplicity, we assume p_a is supported in $[0, 1]$;
- In the bandit setting, only z_{t,a_t} is revealed to the player at time step t;

After n rounds of interactions, the player plays an action sequence of $(a_1, \dots, a_n)$ which is itself random due to the randomness from the player's strategy (and the reward vectors $\{z_t\}_{t\in[n]}$, if the player policy adapts to the observed data $\{z_{t,a_t}\}_{t\in[n]}$. The goal of the player is to design an arm selection strategy such that we have "small" pseudo-regret R_n after n rounds (in hindsight), where

$$R_n := n\mu_* - \sum_{t=1}^{n} \mu_{a_t},$$

where $\mu_* := \max_{a\in[k]} \mu_a$. The pseudo-regret R_n measures the "goodness" of the action sequence $(a_1, \dots, a_n)$ against the best action sequence $(a_*, \dots, a_*)$, where $a_* \in \arg\max_{a\in[k]} \mu_a$. Since R_n is random, there are typically two ways to evaluate R_n.

Expected pseudo-regret. The first way to evaluate R_n is through its expected value, namely expected pseudo-regret

$$\mathbb{E}[R_n] = n\mu_* - \mathbb{E}\left[\sum_{t=1}^{n} \mu_{a_t}\right],$$

where $\mathbb{E}$ is taken over the randomness of the player policy and the reward vectors.

[0]These notes are partially based on those of Patrick Rebeschini.

High-probability pseudo-regret. A stronger[1] notion of expected pseudo-regret is high-probability regret, which is typically stated in the following form: For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$R_n \le f(n,k,1/\delta),$$

where f is some function in n, k and $1/\delta$. A desirable guarantee we want is that

$$\lim_{n\to\infty} \frac{\mathbb{E}[R_n]}{n} = 0, \text{ or } \lim_{n\to\infty} \frac{f(n,k,1/\delta)}{n} = 0.$$

For any $t \in [n]$ and $a \in [k]$, let us define

$$N_{t,a} := \sum_{i=1}^{t-1} \mathbb{1}\{a = a_i\} \tag{15.1}$$

$$\widehat{\mu}_{t,a} := \begin{cases} \frac{1}{N_{t,a}} \sum_{i=1}^{t-1} z_{i,a_i} \mathbb{1}\{a = a_i\} \text{ if } n_{t,a} > 0 \\ 0 \text{ if } n_{t,a} = 0 \end{cases} \tag{15.2}$$

$$\Delta_a := \mu_* - \mu_a \tag{15.3}$$

$$\Delta_{\min} = \min_{a \neq a_*} \Delta_a \tag{15.4}$$

If $\mu_a = \mu_*, \forall a$, the bandit problem is trivial. Thus, we only consider the case there is at least one sub-optimal arm. Without loss of generality, we assume that there is a unique optimal arm a_*, i.e., $|\arg\max_{a\in[k]} \mu_a| = 1$, i.e., $\Delta_{\min} > 0$.

Lemma 15.1 *We have*

$$\mathbb{E}R_n = \sum_{a\in[k]} \Delta_a \mathbb{E}[N_{n,a}].$$

15.1 Explore-then-Commit

To appreciate the non-trivialness of stochastic MAB problem above, let us first start with a simple exploration strategy, namely Explore-then-Commit. That is, the player explores each action for m times in the first mk rounds and commit to the best estimated action for the rest $n - mk$ rounds (assuming $n > mk$).

Lemma 15.2 *For any $m \in \mathbb{N}_+$ that is constant with respect to n, there exists a bandit instance such that the Explore-then-Commit algorithm suffers the expected regret*

$$\mathbb{E}[R_n] \ge c_1 n + c_2,$$

for some absolute constant $c_1 > 0$ and $c_2 \in \mathbb{R}$ that are independent of n.

Fix any $m \in \mathbb{N}_+$. Consider a bandit instance with two arms (i.e., $k = 2$) where $\mu_1 \in [0,1]$ with $p_1 = \delta_{\mu_1}$ and $p_2 = \text{Bernoulli}(\mu_2)$ where $0 < \mu_2 < \mu_1$. The probability that the explore-then-commit algorithm chooses the sub-optimal arm after its exploration phase is

$$p := \Pr(\widehat{\mu}_{2m,1} < \widehat{\mu}_{2m,2}) = \Pr(m\mu_1 < \text{Binomial}(m,\mu_2)) > 0.$$

[1] in the sense that the guarantee in high-probability pseudo-regret typically implies the guarantee in expected pseudo-regret

Algorithm 1 Explore-then-Commit

Input: $m \in \mathbb{N}$, number of arms k, number of rounds n
for $t = 1, \ldots, mk$ **do**
    Play $a_t = t (\text{mode } k) + 1$
    Observe z_{t,a_t}
end for
for $t = mk + 1, \ldots, n$ **do**
    Play $a_t \in \arg\max_{a \in [k]} \widehat{\mu}_{mk,a}$, where $\widehat{\mu}_{mk,a}$ is defined in Eq. (15.3)
    Observe z_{t,a_t}
end for
Output: $\{(a_t, z_{t,a_t})\}_{t \in [n]}$

Thus, we have

$$\begin{aligned}
\mathbb{E}[R_n] &= \mathbb{E}[R_n \mathbb{1}\{\widehat{\mu}_{2m,1} < \widehat{\mu}_{2m,2}\}] + \mathbb{E}[R_n \mathbb{1}\{\widehat{\mu}_{2m,1} \geq \widehat{\mu}_{2m,2}\}] \\
&\geq \mathbb{E}[R_n \mathbb{1}\{\widehat{\mu}_{2m,1} < \widehat{\mu}_{2m,2}\}] \\
&= m(\mu_1 - \mu_2) + (n - 2m)(\mu_1 - \mu_2) \\
&= (n - m)(\mu_1 - \mu_2).
\end{aligned}$$

■

15.2 Epsilon Greedy

We see that Explore-then-Commit suffers from linear (expected) pseudo-regret due to too little explorations. Now let us consider a more "intricate" algorithm that randomly explores an action frequently while also adaptively estimating the mean reward for each action.

Algorithm 2 Epsilon Greedy

Input: number of arms k, number of rounds n, $\epsilon \in [0, 1]$
for $t = 1, \ldots, k$ **do**
    Play $a_t = t$
    Observe z_{t,a_t}
end for
for $t = k + 1, \ldots, n$ **do**
    Play $a_t \sim \text{Uniform}([k])$ with probability ϵ and play $a_t \in \arg\max_{a \in [k]} \widehat{\mu}_{t,a}$ with probability $1 - \epsilon$, where $\widehat{\mu}_{t,a}$ is defined in Eq. (15.3)
    Observe z_{t,a_t}
end for
Output: $\{(a_t, z_{t,a_t})\}_{t \in [n]}$

Lemma 15.3 *For any $\epsilon > 0$, and $n > k$, greedy(ϵ) yields (for any bandit instance),*

$$\mathbb{E}R_n \geq c_1 n + c_2$$

for some absolute constants $c_1 > 0$, $c_2 \in \mathbb{R}$ that are independent of n.

Proof

For any $a \in [k]$, we have

$$N_{n,a} \geq 1 + \frac{\epsilon}{k}(n-k).$$

Thus, we have

$$\mathbb{E}R_n = \sum_{a \in [k]} \mathbb{E}[N_{n,a}\Delta_a] \geq (1 + \frac{\epsilon}{k}(n-k))\Delta_{\min} = c_1 n + c_2.$$

■

The intuition for this failure of ϵ-Greedy is that *the random exploration does not use any information from the history*. Note that for $\epsilon = 0$, the proof above could not show if the greedy policy yields a linear regret. Even in this case, ϵ-Greedy still yields a linear regret.

Lemma 15.4 *There exists an bandit instance such that ϵ-Greedy for $\epsilon = 0$ yields a linear regret.*

Consider a bandit instance with two arms ($k = 2$), where $p_1 = \text{Bernoulli}(\mu_1)$ and $p_2 = \text{Bernoulli}(\mu_2)$ where $0 \leq \mu_2 \leq \mu_1 < 1$. Consider the event $E = \{a_1 = 1, z_{1,a_1} = 0, a_2 = 2, z_{2,a_2} = 1\}$. Under this event E, Greedy($\epsilon = 0$) always commits to playing $a_t = 2$ for all $t \geq 3$ since $\widehat{\mu}_{t,1} = 0, \forall t$. In addition, we have

$$\Pr(E) = (1-\mu_1)\mu_2 > 0.$$

Thus, we have

$$\begin{aligned}
\mathbb{E}[R_n] &= \mathbb{E}[R_n|E]\Pr(E) + \mathbb{E}[R_n|E^c]\Pr(E^c) \\
&\geq \mathbb{E}[R_n|E]\Pr(E) \\
&= (n-1)\Delta_{\min}(1-\mu_1)\mu_2.
\end{aligned}$$

■

15.3 Upper Confidence Bound (UCB)

The algorithms we have considered so far all yield linear regrets. These results stem from that the explorations in these algorithms are not adaptive, in the sense that it completely ignores the collected data to determine where to explore next. Now we consider the celebrated algorithm that has adaptive exploration mechanism, namely Upper Confidence Bounds (UCB). UCB is inspired from the optimism principle in the face of uncertainty (OFUL) that aims at overestimating the mean reward of each action and then selects the next action with the highest mean "optimistic" estimate. By intuition, OFUL encourages to select an action that has not been visited frequently before. In concrete, UCB constructs an optimistic mean estimate for an action using its mean estimate plus some confidence term that inversely relates the frequency at which that action was selected in the previous rounds. Specifically, at round t, UCB selects $a_t \in \arg\max_{a \in [k]} U_{t,a}$, where

$$U_{t,a} = \widehat{\mu}_{t,a} + \sqrt{\frac{\beta_t}{N_{t,a}}},$$

and $\beta_t > 0$ are some confidence parameters to be tuned later.

Algorithm 3 UCB(β)

Input: Confidence parameters $\beta_t > 0$
for $t = 1, \ldots, k$ **do**
    Play $a_t = t$ and receive z_{t,a_t}
end for
for $t = k+1, \ldots, n$ **do**
    Play $a_t = \arg\max_{a\in[k]} U_{t,a}$ and receive z_{t,a_t}
end for
Output: $\{(a_t, z_{t,a_t})\}_{t\in[n]}$

Lemma 15.5 *For any $\delta > 0$, if we set*

$$\beta_t = \beta_\delta = \sqrt{0.5\log(4(n-k)/\delta)}, \forall t \in [n],$$

then the expected regret of UCB is

$$\mathbb{E}R_n \le 2\log(4(n-k)/\delta) \sum_{a\neq a_*} \frac{1}{\Delta_a} + n\delta \sum_{a\in[k]} \Delta_a.$$

Remark 1 *Lemma 15.5 holds for any $\delta > 0$. By simply choosing $\delta = 1/n$, Lemma 15.5 shows that the expected (pseudo)-regret of UCB with a proper choice of β_t grows only log-polynomially with n and depends on the sub-optimality gaps Δ_a.*

Proof of Lemma 15.5 The key idea is to show that for any non-optimal action should not be picked too frequently by UCB. In fact, we will roughly show that $N_{n,a} \lesssim \frac{1}{\Delta_a^2}$ for any non-optimal action a. Let $L_{t,a} = \widehat{\mu}_{t,a} - \sqrt{\frac{\beta_t}{N_{t,a}}}$. Informally, $\mu_a \in [L_{t,a}, U_{t,a}]$. When a sub-optimal action a has been played in the number of times that $N_{n,a} \gtrsim \frac{1}{\Delta_a^2}$, roughly, we have

$$U_{t,a} = L_{t,a} + 2\sqrt{\frac{\beta_t}{N_{t,a}}} \lesssim \mu_a + \Delta_a = \mu_* \le U_{t,a_*}.$$

thus, a could not be selected by UCB at time t.

Now, we formally prove the regret of UCB based on the intuition above. By Hoeffding's inequality, for any $\delta > 0$ and any $a \in [k]$, we have

$$\Pr\left(|\mu_a - \widehat{\mu}_{t,a}| \le \sqrt{\frac{\log(2/\delta)}{2N_{t,a}}}\middle| N_{t,a}\right) \ge 1 - \delta$$

Thus, for any $t \in [k+1, n]$ we have

$$\begin{aligned}\Pr\left(|\mu_a - \widehat{\mu}_{t,a}| \le \sqrt{\frac{\log(2/\delta)}{2N_{t,a}}}\right) &= \sum_{i=1}^{n} \Pr\left(|\mu_a - \widehat{\mu}_{t,a}| \le \sqrt{\frac{\log(2/\delta)}{2N_{t,a}}}\middle| N_{t,a} = i\right) \Pr(N_{t,a} = i) \\ &\ge \sum_{i=1}^{n} (1-\delta)\Pr(N_{t,a} = i)\end{aligned}$$

$$= 1 - \delta.$$

Using the union bound, for any $a \in [k]$, we have

$$\Pr\left(|\mu_a - \widehat{\mu}_{t,a}| \leq \sqrt{\frac{\log(2(n-k)/\delta)}{2N_{t,a}}}, \forall t \in [k+1, n]\right) \geq 1 - \delta.$$

Fix any sub-optimal action a. Let us consider the event that

$$E_a := \{\mu_a \in [L_{t,a}, U_{t,a}], \forall t \in [k+1, n]\} \cap \{\mu_* \in [L_{t,a^*}, U_{t,a^*}], \forall t \in [k+1, n]\}, \tag{15.5}$$

where

$$L_{t,a} := \widehat{\mu}_{t,a} - \sqrt{\frac{\log(4(n-k)/\delta)}{2N_{t,a}}},$$
$$U_{t,a} := \widehat{\mu}_{t,a} + \sqrt{\frac{\log(4(n-k)/\delta)}{2N_{t,a}}}.$$

We have that $\Pr(E_a) \geq 1 - \delta, \forall a$. Let n_a be the large iteration $t \in [n]$ that action a is played at iteration t. Since a is played at iteration n_a, we must have $U_{n_a,a} > U_{n_a,a_*}$. This implies that, under event E_a, we have

$$\begin{aligned}
\mu_a &\geq L_{n_a,a} \\
&= U_{n_a,a} - 2\sqrt{\frac{\log(4(n-k)/\delta)}{2N_{n_a,a}}} \\
&\geq U_{n_a,a^*} - 2\sqrt{\frac{\log(4(n-k)/\delta)}{2N_{n_a,a}}} \\
&\geq \mu_* - 2\sqrt{\frac{\log(4(n-k)/\delta)}{2N_{n_a,a}}}.
\end{aligned}$$

Thus, we have

$$N_{n_a,a} \leq \frac{2\log(4(n-k)/\delta)}{\Delta_a^2}.$$

Since n_a is the last iteration that a is played, we have $N_{n,a} = N_{n_a,a} \leq \frac{2\log(4(n-k)/\delta)}{\Delta_a^2}$. Thus, we have

$$\Delta_a \mathbb{E}[N_{n,a}] = \Delta_a \mathbb{E}[N_{n,a} \mathbb{1}\{E_a\}] + \Delta_a \mathbb{E}[N_{n,a} \mathbb{1}\{E_a^c\}] \leq \frac{2\log(4(n-k)/\delta)}{\Delta_a} + n\Delta_a\delta. \tag{15.6}$$

■

Lemma 15.5 yields a bound that depends on the sub-optimality gaps Δ_a which are instance-dependent quantities. It is possible to modify the proof of Lemma 15.5 to obtain a "minimax" bound that holds regardless any bandit instance.

Lemma 15.6 *If we set β_t as in Lemma 15.5 and $\delta = 1/n$, we have*

$$\mathbb{E}R_n \leq \sqrt{2nk\log(4n(n-k))} + k.$$

Proof of Lemma 15.6 For any $\delta > 0$, $\epsilon > 0$, we have

$$\begin{aligned}
\mathbb{E}R_n &= \sum_{a\in[k]} \Delta_a \mathbb{E}[N_{n,a}] (\text{Lemma 15.1}) \\
&= \sum_{a\in[k]} \Delta_a \mathbb{E}[N_{n,a}]\mathbb{1}\{\Delta_a < \epsilon\} + \sum_{a\in[k]} \Delta_a \mathbb{E}[N_{n,a}]\mathbb{1}\{\Delta_a \geq \epsilon\} \\
&\leq \epsilon n + \sum_{a\in[k]} \Delta_a \mathbb{E}[N_{n,a}]\mathbb{1}\{\Delta_a \geq \epsilon\} \\
&= \epsilon n + \sum_{a\in[k]} \Delta_a \mathbb{E}[N_{n,a}\mathbb{1}\{E_a\}]\mathbb{1}\{\Delta_a \geq \epsilon\} + \sum_{a\in[k]} \Delta_a \mathbb{E}[N_{n,a}\mathbb{1}\{E_a^c\}]\mathbb{1}\{\Delta_a \geq \epsilon\} \quad (E_a \text{ defined in Eq. (15.5)}) \\
&\leq \epsilon n + \sum_{a\in[k]} \frac{2\log(4(n-k)/\delta)}{\Delta_a}\mathbb{1}\{\Delta_a \geq \epsilon\} + \sum_{a\in[k]} n\Delta_a\delta\mathbb{1}\{\Delta_a \geq \epsilon\} \quad (\text{Eq. (15.6)}) \\
&\leq \epsilon n + \sum_{a\in[k]} \frac{2\log(4(n-k)/\delta)}{\epsilon}\mathbb{1}\{\Delta_a \geq \epsilon\} + \sum_{a\in[k]} n\delta \quad (\Delta_a \leq 1) \\
&\leq \epsilon n + \frac{2k\log(4(n-k)/\delta)}{\epsilon} + nk\delta.
\end{aligned}$$

Note that the above inequality holds for any $\epsilon > 0$. Picking $\delta = 1/n$ and minimizing the RHS of the above inequality with respect to ϵ yields

$$\mathbb{E}R_n \leq \sqrt{2nk\log(4n(n-k))} + k.$$

■