Algorithms and Uncertainty, Winter 2025/26 Lecture 17 (5 pages)

Stochastic Multi-Armed Bandits

Thomas Kesselheim Last Update: January 7, 2026

So far, we have seen examples of online algorithms, in which we know nothing about the future; it can be arbitrary. We have also seen examples of Markov Decision processes, in which we have precise knowledge of the underlying stochastic process and its involved probabilities. Today, we will move to the space in between. There is an underlying stochastic process but we do not know it. In contrast to online algorithms, we can try to "learn" the involved probability distributions while making decisions.

1 Model

We consider a version of a "multi-armed bandit". There are n arms we can choose from in every time step. Whenever pulling arm i, the reward is an independent draw from some probability distribution. The crux is that we do not know these distributions. The only thing we can do is repeatedly pull arms, observe the reward, and make our decision which arms to pull based on the observed rewards up to this point.

In more detail, there are n arms with initially unknown reward probability distributions $\mathcal{D}_1, \dots, \mathcal{D}_n$ with means $\mu_1, \dots, \mu_n$. In step t, the algorithm chooses an arm I_t and experiences reward $R_t \in [0,1]$ drawn from $\mathcal{D}_{I_t}$. So, the overall expected reward is $\mathbf{E}\left[\sum_{t=1}^{T} R_t\right]$. We assume that $T \geq n$.

If we had perfect knowledge of $\mathcal{D}_1, \dots, \mathcal{D}_n$, the expected reward would be maximized by always choosing arm i^* defined by $\mu_{i^*} = \max_i \mu_i$.[1] The *expected regret* of an algorithm is the difference between this expected reward and the one experienced by the algorithm

$$\text{Regret} = T \cdot \max_i \mu_i - \mathbf{E}\left[\sum_{t=1}^{T} R_t\right] \quad .$$

2 A Simple Explore-Exploit Algorithm

A very simple algorithm works as follows. There are two phases. In the *exploration* phase, we try out each arm exactly k times, so the length of this phase is kn steps. Afterwards, we have k samples from each distribution. From this, we can compute an empirical average $\hat{\mu}_i$ for each arm. If k is large enough, then $\hat{\mu}_i$ should be close to the actual mean μ_i. Therefore, in the *exploitation* phase (of length $T - kn$), we always play the arm that maximizes $\hat{\mu}_i$.

There is a clear trade-off between the exploration and the exploitation. If we set k too small, then $\hat{\mu}_i$ is computed only based on a few samples and can be far from μ_i. If k is too large, the exploration phase takes too long, during which we also play very bad arms.

Theorem 17.1. *Setting $k = \left(\frac{T}{n}\right)^{\frac{2}{3}}$, the expected regret of the simple algorithm is upper-bounded by $O(n^{\frac{1}{3}} T^{\frac{2}{3}} \ln(nT))$.*

To prove this theorem, we will use again Hoeffding's inequality and the union bound. Hoeffding's inequality is a quantitative version of the law of large numbers. It states that we get close to the expectation if we take the average of sufficiently many independent draws from a distribution.

[1] For simplicity we assume that this arm is unique.

Lemma 17.2 (Hoeffding's inequality)**.** *Let $Z_1, \ldots, Z_N$ be independent random variables such that $a_i \leq Z_i \leq b_i$ with probability 1. Let $\bar{Z} = \frac{1}{N}\sum_{i=1}^{N} Z_i$ be their average. Then for all $\gamma \geq 0$*

$$\mathbf{Pr}\left[|\bar{Z} - \mathbf{E}\left[\bar{Z}\right]| \geq \gamma\right] \leq 2\exp\left(-\frac{2N^2\gamma^2}{\sum_{i=1}^{N}(b_i - a_i)^2}\right) \ .$$

The union bound gives an easy upper bound on the probability of a union of events.

Lemma 17.3 (Union Bound)**.** *For any sequence of not necessarily disjoint events $\mathcal{E}_1, \mathcal{E}_2 \ldots$, we have*

$$\mathbf{Pr}\left[\mathcal{E}_1 \cup \mathcal{E}_2 \cup \ldots\right] \leq \mathbf{Pr}\left[\mathcal{E}_1\right] + \mathbf{Pr}\left[\mathcal{E}_2\right] + \ldots \ .$$

Proof of Theorem 17.1. By Hoeffding's inequality, we get that for all arms i and all $\gamma > 0$

$$\mathbf{Pr}\left[|\hat{\mu}_i - \mu_i| \geq \gamma\right] \leq 2\exp\left(-\frac{2k^2\gamma^2}{k(1-0)^2}\right) = 2\exp(-2k\gamma^2) \ .$$

In combination with the union bound, we get

$$\mathbf{Pr}\left[\exists i : |\hat{\mu}_i - \mu_i| \geq \gamma\right] \leq \sum_{i=1}^{n} \mathbf{Pr}\left[|\hat{\mu}_i - \mu_i| \geq \gamma\right] \leq 2n\exp(-2k\gamma^2) \ .$$

If $|\hat{\mu}_i - \mu_i| < \gamma$ for all i, this means the following. Consider the arm i that is played during the exploitation phase. We have

$$\mu_i \geq \hat{\mu}_i - \gamma \geq \hat{\mu}_{i^*} - \gamma \geq \mu_{i^*} - 2\gamma \ .$$

Every reward is always non-negative. So, we can lower-bound the overall expected reward by taking into consideration only the rewards from the exploitation phase in the case that $|\hat{\mu}_i - \mu_i| < \gamma$ for all i. Writing ϵ for $2n\exp(-2k\gamma^2)$, we get in combination

$$\begin{aligned}\mathbf{E}\left[\sum_{t=1}^{T} R_t\right] &\geq \mathbf{Pr}\left[\forall i : |\hat{\mu}_i - \mu_i| < \gamma\right](T - nk)(\mu_{i^*} - 2\gamma) \\ &\geq (1-\epsilon)(T - nk)(\mu_{i^*} - 2\gamma) \\ &\geq T\mu_{i^*} - \epsilon T - nk - T2\gamma \ ,\end{aligned}$$

where we also used that $\mu_{i^*} \leq 1$ and $\epsilon \leq 1$.

Therefore, the expected regret will be at most

$$T\mu_{i^*} - \mathbf{E}\left[\sum_{t=1}^{T} R_t\right] \leq \epsilon T + nk + 2\gamma T \ .$$

This bound holds for all k and all γ. Observe that $\gamma = \sqrt{\frac{1}{2k}\ln\frac{2n}{\epsilon}}$. If we now choose $k = \left(\frac{T}{n}\right)^{\frac{2}{3}}$, $\epsilon = \left(\frac{n}{T}\right)^{\frac{1}{3}}$ and the γ that is implied by this choice, the bound holds. □

3 UCB1 Algorithm

One of the major weaknesses of the simple algorithm is that it does not adapt the exploration to its observations: If an arm turns out to be very bad, intuitively it should be clear that it can be ignored after only a few samples.

The UCB1 algorithm is smarter. UCB stands for *upper confidence bound*. This almost explains the entire algorithm. We make no further distinction between exploration and exploitation. In each step, we use the empirical averages $\hat{\mu}_i^{(t)}$ of the arms observed so far. The actual means μ_i are close. They are closer, the more often we pulled the respective arm. Out of the empirical average and the number of times we pulled the arm so far, Q_i, we can compute confidence bounds, in which we expect the μ_i values to lie. We are optimistic and choose the arm with the highest upper bound on μ_i.

In more detail, in the first n steps, we pull each arm once. In any step $t > n$, we let $\hat{\mu}_i^{(t)}$ be the empirical average of arm i seen in steps *before* t. Furthermore, we let $Q_i^{(t)}$ denote the number of times that we pulled arm i *before* step t. In step t, we choose the arm i for which $\hat{\mu}_i^{(t)} + \sqrt{\frac{2\ln T}{Q_i^{(t)}}}$ is highest.

Theorem 17.4. *The expected regret of UCB1 is at most* $4\sqrt{2nT\ln T} + n + 1$.

Below, we will prove the following lemma. It gives exactly the confidence bounds we use in the algorithm.

Lemma 17.5. *We have*

$$\mathbf{Pr}\left[\exists i\ \exists t : |\hat{\mu}_i^{(t)} - \mu_i| \geq \sqrt{\frac{2\ln T}{Q_i^{(t)}}}\right] \leq \frac{1}{T}\ .$$

Based on the lemma, we can bound the expected reward we obtain in a single step of the algorithm.

Lemma 17.6. *The expected reward obtained in step* $t \geq n+1$ *fulfills*

$$\mathbf{E}\left[R_t\right] \geq \mu_{i^*} - \mathbf{E}\left[2\sqrt{\frac{2\ln T}{Q_{I_t}^{(t)}}}\right] - \frac{1}{T}\ .$$

Proof. Let's fix everything that happens before step t. This fixes which arm I_t will be pulled in step t. The expected reward from this pull is μ_{I_t}. Let's suppose that

$$|\hat{\mu}_i^{(t)} - \mu_i| < \sqrt{\frac{2\ln T}{Q_i^{(t)}}} \qquad \text{for all arms } i. \tag{1}$$

Instead of pulling arm I_t we could also pull arm i^*. Therefore, we have

$$\hat{\mu}_{I_t}^{(t)} + \sqrt{\frac{2\ln T}{Q_{I_t}^{(t)}}} \geq \hat{\mu}_{i^*}^{(t)} + \sqrt{\frac{2\ln T}{Q_{i^*}^{(t)}}}\ .$$

In combination, we get

$$\mu_{I_t} \geq \hat{\mu}_{I_t}^{(t)} - \sqrt{\frac{2\ln T}{Q_{I_t}^{(t)}}} \geq \hat{\mu}_{i^*}^{(t)} + \sqrt{\frac{2\ln T}{Q_{i^*}^{(t)}}} - 2\sqrt{\frac{2\ln T}{Q_{I_t}^{(t)}}} \geq \mu_{i^*} - 2\sqrt{\frac{2\ln T}{Q_{I_t}^{(t)}}}\ .$$

This bound only holds if (1) is fulfilled. If (1) is not fulfilled, we still have $\mu_{I_t} \geq 0$. So letting $Z = 1$ if (1) is not fulfilled $Z = 0$ otherwise, we always have

$$\mu_{I_t} \geq \mu_{i^*} - 2\sqrt{\frac{2\ln T}{Q_{I_t}^{(t)}}} - Z \ .$$

Taking the expectation, we obtain $\mathbf{E}[\mu_{I_t}] \geq \mu_{i^*} - \mathbf{E}\left[2\sqrt{\frac{2\ln T}{Q_{I_t}^{(t)}}}\right] - \mathbf{E}[Z]$. Finally, $\mathbf{E}[Z] = \mathbf{Pr}[Z = 1] \leq \frac{1}{T}$ follows from Lemma 17.5. □

Proof of Theorem 17.4. By Lemma 17.6, the expected regret is

$$\text{Regret} = T\mu_{i^*} - \mathbf{E}\left[\sum_{t=1}^{T} R_t\right] \leq n + \sum_{t=n+1}^{T}\left(\mathbf{E}\left[2\sqrt{\frac{2\ln T}{Q_{I_t}^{(t)}}}\right] + \frac{1}{T}\right) = 2\sqrt{2\ln T}\,\mathbf{E}\left[\sum_{t=n+1}^{T}\sqrt{\frac{1}{Q_{I_t}^{(t)}}}\right] + n + 1 \ .$$

It only remains to bound the sum inside the expectation. Note that we can reorder it: For every arm i, sum up the t for which $I_t = i$. Then we get the sequence of terms $\sqrt{\frac{1}{k}}$ for $k = 1, \ldots Q_i - 1$. Furthermore, note that $\sum_{k=1}^{Q_i-1}\sqrt{\frac{1}{k}} \leq \int_0^{Q_i}\sqrt{\frac{1}{x}}\mathrm{d}x = 2\sqrt{Q_i}$. So, we obtain

$$\sum_{t=n+1}^{T}\sqrt{\frac{1}{Q_{I_t}^{(t)}}} = \sum_{i=1}^{n}\sum_{k=1}^{Q_i-1}\sqrt{\frac{1}{k}} \leq \sum_{i=1}^{n} 2\sqrt{Q_i} \ .$$

We could already conclude that this term is no more that $n\sqrt{T}$ because all Q_i are no more than T. However, we even have $\sum_{i=1}^{n} Q_i = T$. Therefore, by concavity of the square-root function, we get

$$\sum_{i=1}^{n}\sqrt{Q_i} = n\frac{1}{n}\sum_{i=1}^{n}\sqrt{Q_i} \leq n\sqrt{\frac{1}{n}\sum_{i=1}^{n} Q_i} = n\sqrt{\frac{1}{n}T} = \sqrt{nT} \ . \qquad \square$$

4 Proof of Lemma 17.5

Proof. The difficulty is that we cannot apply Hoeffding's inequality on a fixed $\hat{\mu}_i^{(t)}$: The value of $\hat{\mu}_i^{(t)}$ depends on how often we have pulled arm i so far and how the rewards turned out to be. If the first rewards were low, the value of $\hat{\mu}_i^{(t')}$ at the earlier time t' is low as well and arm i gets ignored.

Therefore, instead, we will consider the following way to determine all values of $\hat{\mu}_i^{(t)}$. First, we draw T values for each arm i, which we denote by $X_{i,1}, \ldots, X_{i,T}$. The meaning is that the j^{th} draw from arm i has reward $X_{i,j}$. Note that fixing all these random outcomes, this also determines what the algorithm does and consequently also $\hat{\mu}_i^{(t)}$.

Now, in order for there to be a t such that $|\hat{\mu}_i^{(t)} - \mu_i| \geq \sqrt{\frac{\ln T}{Q_i^{(t)}}}$, there has to be a k such that $|\frac{1}{k}\sum_{j=1}^{k} X_{i,j} - \mu_i| \geq \sqrt{\frac{\ln T}{k}}$. On these random variables $X_{i,1}, \ldots, X_{i,k}$, we can apply Hoeffding's inequality and get

$$\mathbf{Pr}\left[\left|\frac{1}{k}\sum_{j=1}^{k} X_{i,j} - \mu_i\right| \geq \sqrt{\frac{2\ln T}{k}}\right] \leq 2\exp\left(-2k\frac{2\ln T}{k}\right) = \frac{2}{T^4} \ .$$

The event that there are an i and k for which $\left|\frac{1}{k}\sum_{j=1}^{k} X_{i,j} - \mu_i\right| \geq \sqrt{\frac{\ln T}{k}}$ is again a union of events for which we can apply the union bound. Therefore

$$\mathbf{Pr}\left[\exists i\ \exists k : \left|\frac{1}{k}\sum_{j=1}^{k} X_{i,j} - \mu_i\right| \geq \sqrt{\frac{2\ln T}{k}}\right] \leq \sum_{i=1}^{n}\sum_{k=1}^{T} \mathbf{Pr}\left[\left|\frac{1}{k}\sum_{j=1}^{k} X_{i,j} - \mu_i\right| \geq \sqrt{\frac{2\ln T}{k}}\right] \leq nT\frac{2}{T^4} \leq \frac{1}{T}\ .$$

□

Reference

Peter Auer, Nicolò Cesa-Bianchi, Paul Fischer: Finite-time Analysis of the Multiarmed Bandit Problem. Mach. Learn. 47(2-3): 235-256 (2002)