**Master MVA: Reinforcement Learning** **Lecture: 4**

# Multi-arm Bandit Framework

Lecturer: *Alessandro Lazaric* *http://researchers.lille.inria.fr/∼lazaric/Webpage/Teaching.html*

**Objectives of the lecture**

1. **Understand:** The multi-armed bandit problem and its extensions.
2. **Use:** UCB, Exp3, and improvements.

## 1 The Stochastic Multi-arm Bandit Problem

We consider $K$ arms (i.e., actions, options) characterized by $K$ unknown distributions $(\nu_k)_{1\leq k\leq K}$ bounded in $[0,1]$. At each step $t$, the learner selects an arm $I_t \in \{1,\dots,K\}$ and observes a reward $x_t \sim \nu_{I_t}$, which is an independent sample drawn from the distribution corresponding to the chosen arm, i.e., $\nu_{I_t}$. The objective of the learner is to maximize the *expected* sum of rewards over time.

Let $\mu_k = \mathbb{E}_{X\sim\nu_k}[X]$ be the expectation of each arm and $\mu^* = \max_k \mu_k$ the expected value of the optimal arm (in expectation). If the learner knew the distribution, it would select the best arm $k^* = \arg\max_k \mu_k$ at each time step, thus obtaining an average reward of $\mu^*$. Since the distributions are unknown, the learner needs to explore all the different arms to collect information (*exploration*) which can later be used to act optimally (*exploitation*). This leads to the so-called **exploration-exploitation dilemma**.

In order to evaluate the performance of a given strategy, we define at which speed the average reward obtained by the strategy converges to the average optimal reward. We introduce the notion of regret as follows.

**Definition 1.** *Given a time horizon of $n$ steps, a given strategy which observes the sequence of rewards $x_{tt}$ suffers from a* **cumulative regret***:*

$$R_n = n\mu^* - \sum_{t=1}^{n} x_t,$$

*where $x_t$ is an i.i.d. realization from the distribution $\nu_{I_t}$ of the arm $I_t$ chosen by the strategy at time $t$.*

The regret measures the difference between the (expected) cumulative reward that would be obtained by repeatedly pulling the optimal arm and the reward accumulated by the strategy.

In particular, we study the regret in expectation, i.e., $\mathbb{E}R_n$, which can be written as

$$\mathbb{E}R_n = n\mu^* - \mathbb{E}\sum_{t=1}^{n} \mu_{I_t} = \mathbb{E}\sum_{k=1}^{K} T_k(n)(\mu^* - \mu_k) = \mathbb{E}\sum_{k=1}^{K} T_k(n)\Delta_k,$$

where $\Delta_k = \mu^* - \mu_k$ is the gap between the optimal arm and arm $k$ and $T_k(n) = \sum_{t=1}^{n} \mathbb{I}\{I_t = k\}$ is the number of times arm $k$ has been pulled until step $n$. Then a good leaner should pull the sub-optimal arms as rarely as possible depending on their gaps.

## 1.1 The Upper-Confidence Bound (UCB) Algorithm

*Algorithm Definition* 1. At each time instant $t$, the UCB [Auer et. al, 2002] strategy pulls the arm

$$I_t = \arg\max_k B_{t,T_k(t-1)}(k), \text{ with } B_{t,s}(k) = \hat{\mu}_{k,s} + \sqrt{\frac{3\log t}{2s}},$$

where $\hat{\mu}_{k,s} = \frac{1}{s}\sum_{i=1}^{s} x_{k,i}$ is the empirical mean of the rewards observed by pulling arm $k$ (i.e., $x_{k,i}$ is the $i$-th reward received from arm $k$).

The UCB strategy follows the celebrated **optimism in face of uncertainty** principle. In fact, the $B_{t,T_k(t-1)}(k)$ value is a high-probability upper bound on the expected value $\mu_k$. Thus, UCB selects the arm which has the best value if all the arms had the best possible value *compatible* with the observations obtained so far.

In fact, from Chernoff-Hoeffding inequality we have that for any sequence of $s$ i.i.d. random variables $X_i \in [0,1]$ with common mean $\mu = \mathbb{E}X_i$ we have that

$$\mathbb{P}\Big(\frac{1}{s}\sum_{i=1}^{s} X_i - \mu \geq \epsilon\Big) \leq e^{-2s\epsilon^2}, \quad \text{and} \quad \mathbb{P}\Big(\frac{1}{s}\sum_{i=1}^{s} X_i - \mu \leq -\epsilon\Big) \leq e^{-2s\epsilon^2}. \tag{1}$$

Thus for any fixed $1 \leq s \leq t$

$$\mathbb{P}\Big(\hat{\mu}_{k,s} + \sqrt{\frac{3\log t}{2s}} \leq \mu_k\Big) \leq e^{-3\log(t)} = t^{-3}. \tag{2}$$

And

$$\mathbb{P}\Big(\hat{\mu}_{k,s} - \sqrt{\frac{3\log t}{2s}} \geq \mu_k\Big) \leq e^{-3\log(t)} = t^{-3}. \tag{3}$$

*Proposition* 1. Any sub-optimal arm $k \neq k^*$ is pulled by UCB at most

$$\mathbb{E}T_k(n) \leq 6\frac{\log n}{\Delta_k^2} + \frac{\pi^2}{3} + 1$$

times. Thus, the corresponding regret is bounded as

$$\mathbb{E}R_n = \sum_k \Delta_k \mathbb{E}T_k(n) \leq 6\sum_{k:\Delta_k>0} \frac{\log n}{\Delta_k} + K\big(\frac{\pi^2}{3} + 1\big).$$

This result implies that UCB has a cumulative regret which grows as $\log(n)$.

*Proof.* Let's start with an intuitive argument. Let suppose that at time $t$ the empirical averages of all the arms are indeed contained in their confidence intervals, that is

$$\mu_k - \sqrt{\frac{3\log t}{2s}} \overset{(a)}{\leq} \hat{\mu}_{k,s} \overset{(b)}{\leq} \mu_k + \sqrt{\frac{3\log t}{2s}}. \tag{4}$$

where $s = T_k(t-1)$. Let $k$ be any suboptimal arm and $k^*$ an optimal arm. If arm $k$ is pulled at time $t$, then by definition of the algorithm it means that $B_{t,T_k(t-1)}(k) \geq B_{t,T_{k^*}(t-1)}(k^*)$, which corresponds to

$$\hat{\mu}_{k,s} + \sqrt{\frac{3\log t}{2s}} \geq \hat{\mu}_{k^*,s^*} + \sqrt{\frac{3\log t}{2s^*}}, \tag{5}$$

where $s = T_k(t-1)$ and $s^* = T_{k^*}(t-1)$. Thus according to(4) we obtain

$$\mu_k + 2\sqrt{\frac{3\log t}{2s}} \geq \mu^*,$$

which corresponds to an upper bound on the number of pulls

$$s \leq \frac{6\log t}{\Delta_k^2}.$$

More in detail, for any positive integer $u$ we have that

$$\begin{aligned} T_k(n) &\leq u + \sum_{t=u+1}^{n} \mathbb{I}\{I_t = k; T_k(t) > u\} \\ &\leq u + \sum_{t=u+1}^{n} \mathbb{I}\{\exists s : u < s \leq t, \exists s^* : 1 \leq s^* \leq t, B_{t,s}(k) \geq B_{t,s^*}(k^*)\} \end{aligned} \tag{6}$$

Following the previous reasoning, the event $\{B_{t,s}(k) \geq B_{t,s^*}(k^*)\}$ (i.e, (5)) implies that $s \leq \frac{6\log t}{\Delta_k^2}$ or that either one of the two inequalities (a) or (b) in (4) are not true. If we choose $u = \frac{8\log(n)}{\Delta_k^2} + 1$, then either (a) or (b) is not satisfied. Nonetheless, from (2), inequality (a) is not true with a probability $\leq t^{-3}$, while from (3) inequality (b) is not true with a probability $\leq t^{-3}$.

Thus, taking the expectation on both sides of (6),

$$\begin{aligned} \mathbb{E}[T_k(n)] &\leq \frac{6\log(n)}{\Delta_k^2} + 1 + \sum_{t=u+1}^{n} \Big[ \sum_{s=u+1}^{t} t^{-3} + \sum_{s=1}^{t} t^{-3} \Big] \\ &\leq \frac{6\log(n)}{\Delta_k^2} + \frac{\pi^2}{3} + 1 \end{aligned}$$

□

Beside the previous regret bound, we can also derive the following *distribution independent* bound.

*Proposition* 2. The UCB algorithm has a (uniform) regret of

$$\mathbb{E}R_n \leq \sqrt{Kn(6\log n + \frac{\pi^2}{3} + 1)}$$

*Proof.* From Cauchy-Schwarz,

$$\begin{aligned} \mathbb{E}R_n &= \sum_k \Delta_k \sqrt{\mathbb{E}T_k(n)} \sqrt{\mathbb{E}T_k(n)} \\ &\leq \sqrt{\sum_k \Delta_k^2 \mathbb{E}T_k(n) \sum_k \mathbb{E}T_k(n)} \\ &\leq \sqrt{Kn(6\log n + \frac{\pi^2}{3} + 1)}. \end{aligned}$$

□

## 1.2 Lower Bounds

We also have an asymptotic distribution-dependent lower-bound (for a rather large set of distributions) [Lai et Robbins, 1985]:

$$\limsup_n \frac{\mathbb{E}T_k(n)}{\log n} \geq \frac{1}{KL(\nu_k||\nu^*)},$$

where the Kullback-Leibler distance is $KL(\nu||\nu') = \int d\nu \log(d\nu/d\nu')$. Then $\mathbb{E}R_n = \Omega(\log n)$.

We also have a non-asymptotic distribution-independent lower-bound (see e.g., [Cesa-Bianchi et Lugosi, Prediction, Learning, and Games, 2006]):

$$\inf_{\text{alg}} \sup_{\text{prob}} R_n = \Omega(\sqrt{nK}).$$

## 1.3 Improvements

- We can use the empirical variance to refine the precision of the confidence intervals. See [Audibert, Munos, Szepesvari, Use of variance estimation in the multi-armed bandit problem, 2008].
- We can use the whole empirical distribution instead of only the empirical average, obtaining KL-UCB algorithms ([Garivier, Capp, The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond, 2011] and [Maillard, Munos, Stoltz, Finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences, 2011]).
- Improved minimax bounds [Audibert, Bubeck, Minimax Policies for Adversarial and Stochastic Bandits, 2009] matching the lower bound $\sqrt{Kn}$.

## 1.4 Extensions

There are a large number of extensions to the problem of stochastic bandit with $K$ arms

- **Bandit in MDP** [Jaksch, Ortner, Auer. Near-optimal regret bounds for reinforcement learning, 2010]. A UCB-based exploration-exploitation strategy in an MDP.
- **Contextual bandits**. A each step $t$, the learner observes a context $x_t \in X$ and take a decision $a_t \in A$. The reward is a function of $a_t$ and $x_t$. The regret is measured w.r.t. a class of strategies $\pi : X \to A$.
- **Bandit with a countable set of arms**. [Wang, Audibert, Munos, Algorithms for inifinitely many-armed bandits, 2008]. Each new arm has a probability $\epsilon^\beta$ of being $\epsilon$-optimal. Then the learner has to trade-off exploration - exploitation - discovery.
- **Linear bandit** [Dani, Hayes, Kakade, Stochastic Linear Optimization under Bandit Feedback, 2008] The learner selects an arm $x_t \in X \subset \mathbb{R}^d$ and the reward is a linear combination $r_t = x_t \cdot \alpha$, where $\alpha \in \mathbb{R}^d$ is an unknown parameter vector. The regret is measure w.r.t. $\max_{x \in X} x \cdot \alpha$.
- **Bandits in a metric space** [Kleinberg, Slivkins, Upfal, Multi-armed bandits in metric spaces, 2008], [Bubeck, Munos, Stoltz, Szepesvari, Online optimization in X-armed bandits, 2008]. The learner chooses an arm $x_t \in X$ in a metric space. The expected reward $f(x_t)$ is assumed to be Lipschitz. The regret is measured w.r.t. $\sup_{x \in X} f(x)$. This is an online optimization problem.

- **Hierarchical bandits** Algorithms UCT [Kocsis et Szepesvari. Bandit based monte-carlo planning., 2006], BAST [Coquelin et Munos, Bandit algorithms for tree search, 2007], HOO [Bubeck, Munos, Stoltz, Szepesvari, Online optimization in X-armed bandits, 2008]. Application to the game of Go (MoGo) [Gelly, Wang, Munos, Teytaud. Modification of UCT with patterns in monte-carlo go, 2006]. Use of the bandit algorithms in a hierarchical structure in order to search in tree structures.
- **Optimistic planning** Use of the optimism in face of uncertainty principle for planning. See [Hren et Munos, Optimistic planning for deterministic systems, 2008], [Bubeck et Munos, Open Loop Optimistic Planning, 2010], [Busoniu, Munos, De Schutter, Babuska, Optimistic planning for sparsely stochastic systems, 2011].

# A Concentration Inequalities

*Proposition* 3 (Chernoff-Hoeffding Inequality). Let $X_i \in [a_i, b_i]$ be $n$ independent random variables with mean $\mu_i = \mathbb{E}X_i$. Then

$$\mathbb{P}\big(\big|\sum_{i=1}^{n} X_i - \mu_i\big| \geq \epsilon\big) \leq 2\exp\Big(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\Big). \tag{7}$$

*Proof.* We have

$$\begin{aligned}
\mathbb{P}(\sum_{i=1}^{n} X_i - \mu_i \geq \epsilon) &= \mathbb{P}(e^{s\sum_{i=1}^{n} X_i - \mu_i} \geq e^{s\epsilon}) \\
&\leq e^{-s\epsilon}\mathbb{E}[e^{s\sum_{i=1}^{n} X_i - \mu_i}], \quad \text{Markov inequality} \\
&= e^{-s\epsilon}\prod_{i=1}^{n}\mathbb{E}[e^{s(X_i - \mu_i)}], \quad \text{independent random variables} \\
&\leq e^{-s\epsilon}\prod_{i=1}^{n} e^{s^2(b_i - a_i)^2/8}, \quad \text{Hoeffding inequality} \\
&= e^{-s\epsilon + s^2\sum_{i=1}^{n}(b_i - a_i)^2/8}
\end{aligned}$$

If we choose $s = 4\epsilon/\sum_{i=1}^{n}(b_i - a_i)^2$, then $\mathbb{P}\big(\sum_{i=1}^{n} X_i - \mu_i \geq \epsilon\big) \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$. Similar computation for $\mathbb{P}\big(\sum_{i=1}^{n} X_i - \mu_i \leq -\epsilon\big)$ leads to the result in eq. (7). □