Stochastic bandits (Part 1)

Pierre Gaillard

Reminder from last lectures

We recall the setting of online convex optimization in Setting 1.

At each time step $t = 1, \dots, T$
- the player chooses an action $\theta_t \in \Theta$ (compact decision set);
- the environment chooses a loss function $f_t : \Theta \to [0,1]$;
- the player suffers loss $f_t(\theta_t)$ and observes
  - the losses of every actions: $f_t(\theta)$ for all $\theta \in \Theta$ → full-information feedback
  - the loss of the chosen action only: $f_t(\theta_t)$ → bandit feedback.

The goal of the player is to minimize his cumulative regret:

$$R_T \overset{\text{def}}{=} \sum_{t=1}^{T} f_t(\theta_t) - \inf_{\theta \in \Theta} \sum_{t=1}^{T} f_t(\theta)\,.$$

Setting 1: Setting of an online learning problem

During the last few lectures, we have reviewed the framework for comprehensive information. We designed algorithms to minimize regret for the different decision spaces Θ and loss assumptions f_t. Most of the algorithms were based on variations in the exponentially weighted average forecaster or online gradient descent. We also found some links with game theory, including the Blackwell approach, two-player zero-sum games, and calibration.

In the next lectures, we will consider the bandit setting, when the player only observes the performance of $f_t(\theta_t)$ but not $f_t(\theta)$ for $\theta \neq \theta_t$. We will start by providing fundamental results for stochastic bandits with finitely many actions, also called K-armed bandits which basically corresponds to $\Theta = \{1, \dots, K\}$ and i.i.d. loss functions f_t. For a thorough introduction to stochastic bandits we refer the interested student to the monographs Bubeck et al. [2012] or Lattimore and Szepesvári [2019].

1 Setting: stochastic bandit with finitely many actions

We state here the setting of stochastic bandits with finitely many actions (also called multi-armed bandit) and fix notations that we will use.

A multi-armed bandit problem is a sequential decision problem defined by a finite set of actions $\Theta \overset{\text{def}}{=} \{1, \dots, K\}$ also called *arms*. We assume that there are K unknown sequences $X_{i,1}, X_{i,2}, \dots$ of rewards in $[0,1]$ associated with each arm $i = 1, \dots, K$. At each round, the player makes a decision by pulling an arm k_t in Θ and observes the corresponding reward[1] $X_{k_t,t}$. The objective of the player is to minimize his cumulative

[1] In the bandit community, it is more common to consider rewards rather than losses.

regret:

$$R_T \overset{\text{def}}{=} \max_{k=1,\dots,K} \sum_{t=1}^{T} X_{k,t} - \sum_{t=1}^{T} X_{k_t,t}\,.$$

In stochastic bandits, we generally assume the sequences to be i.i.d. Each arm $k = 1, \dots, K$ is associated an unknown probability distribution ν_k over $[0, 1]$ and $X_{k,t} \sim \nu_k$. We also denote

$$\mu_k \overset{\text{def}}{=} \mathbb{E}[X_{k,t}], \qquad \text{and} \qquad \mu^* \in \operatorname*{arg\,max}_{k=1,\dots,K}\{\mu_k\}\,.$$

The player aims at finding the arm with the highest mean reward μ_k as quickly as possible. The setting is summarized in Setting 2. Note that we retrieve the setting of online optimization (Setting 1) with the notation $X_{k,t} \overset{\text{def}}{=} 1 - f_t(k)$ with i.d.d. loss functions.

Unknown parameters: K probability distributions $\nu_1, \dots, \nu_K$ on $[0, 1]$

At each time step $t = 1, \dots, T$

– the player chooses an action $k_t \in \Theta = \{1, \dots, K\}$;
– given k_t, the environment draws the reward $X_{k_t,t} \sim \nu_{k_t}$;
– the player only observes the feedback $X_{k_t,t}$.

Setting 2: Setting of stochastic bandit with finitely many actions

Multi-armed bandits have several concrete historical applications in a variety of fields, including ad placement, clinical trials, source routing or game AI. The name bandit refers to the “slot machine” in casinos, and the bandit problem corresponds to a player that inserts coins into different machines and tries to maximize his payoff. The finite arms bandit settings we consider are simple enough to be analyzed and the algorithms can often be generalized to more realistic settings including for example contextual bandits.

Remark. Assume that all arms $\nu_k \sim \mathcal{B}(1/2)$ for $k = 1, \dots, K$. Then, $\mathbb{E}[X_{k,t}] = 1/2$ and $\mathbb{E}[X_{k_t,t}] = 1/2$. But because of fluctuations of random walks, the expected magnitude of the maximum rewards is of order

$$\mathbb{E}\left[\max_{k=1,\dots,K} \sum_{k=1}^{T} X_{k,n}\right] \approx \sqrt{T \log K}\,.$$

Therefore, in this case though all arms are optimal, the expected regret is of order $\sqrt{T \log K}$. We will thus consider a more quantity in the stochastic framework called the pseudo-regret which corresponds to competing with the best action in expectation, rather than the optimal action on the sequence of realized rewards.

Definition 1 (Pseudo-regret)**.** *The pseudo-regret is defined as*

$$\bar{R}_T \overset{\textit{def}}{=} T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{k_t}\right],$$

where we recall $\mu_k = \mathbb{E}[X_{k,t}]$.

Remark that the pseudo-regret is upper-bounded by the expected regret $\bar{R}_T \le \mathbb{E}\big[R_T\big]$. It is thus harder to design algorithms for the true regret but we will focus here on the pseudo-regret.

Useful notation In the following, we will denote by $\widehat{\mu}_k(s)$ the empirical mean of rewards obtained after pulling arm k s times. Let us also denote for all arms $k = 1, \dots, K$ by

$$\Delta_k \overset{\text{def}}{=} \mu^* - \mu_k \qquad \text{and} \qquad N_k(t) \overset{\text{def}}{=} \sum_{s=1}^{t} \mathbb{1}_{k_t = k}\,,$$

respectively the suboptimal gap of arm k and the number of times the arm k was pulled by the player before time t. Then, the pseudo-regret can be rewritten

$$\bar{R}_T = \bigg(\sum_{k=1}^{K} \mathbb{E}\big[N_k(t)\big] \bigg) \mu^* - \mathbb{E}\bigg(\sum_{k=1}^{K} N_k(t) \mu_k \bigg) = \sum_{k=1}^{K} \Delta_k \mathbb{E}\big[N_k(t)\big]\,. \tag{1}$$

We recall Hoeffding's inequality that will be used in the proofs. We will often use Azuma-Hoeffding's inequality which is a generalization of Hoeffding's inequality to martingals.

Proposition 1 (Hoeffding's Inequality)**.** *If $X_1, \dots, X_n$ are independent random variables almost surely in $[a, b]$ then for all $\delta \in (0, 1)$ we have*

$$\mathbb{P}\left\{ \sum_{t=1}^{n} X_k - \mathbb{E}\bigg[\sum_{t=1}^{n} X_k \bigg] \geq (b - a) \sqrt{\frac{n}{2} \log \frac{1}{\delta}} \right\} \leq \delta\,,$$

or equivalently for all $\varepsilon > 0$

$$\mathbb{P}\left\{ \sum_{t=1}^{n} X_k - \mathbb{E}\bigg[\sum_{t=1}^{n} X_k \bigg] \geq \varepsilon \right\} \leq \exp\bigg(- \frac{2\varepsilon^2}{n(b-a)^2} \bigg)\,.$$

2 Explore-Then-Commit (ETC)

Contrary to the full information we examined earlier, the player only observes the rewards of the chosen actions. He must therefore make a trade-off between exploration and exploitation. The first bandit algorithm that we consider is Explore Then Commit (ETC). It consists in first performing an exploration phase of mK length in which each arm is pulled $m \geq 1$ times. Then it exploits by pulling the arm with the best empirical reward for the remaining rounds. It is summarized in Algorithm 1.

```
Parameter: $m \geq 1$.
1. Exploration
    – For rounds $t = 1, \dots, mK$ explore by drawing each arm $m$ times.
    – Compute for each arm $k$ its empirical mean of rewards obtained by pulling arm $k$ $m$ times
          $\widehat{\mu}_k(m) = \frac{1}{m} \sum_{s=1}^{Km} \mathbb{1}_{k_t = k} X_{k,t}\,.$
2. Exploitation: keep playing the best arm $\arg\max_k \widehat{\mu}_k(m)$ for the remaining rounds $t = mK + 1, \dots, T$.
```

Algorithm 1: Explore-Then-Commit (ETC)

Theorem 1 (Thm 6.1, Lattimore and Szepesvári [2019])**.** *If $1 \leq m \leq T/K$ then*

$$\bar{R}_T \leq m \sum_{k=1}^{K} \Delta_k + (T - mK) \sum_{k=1}^{K} \Delta_k \exp\big(- m\Delta_k^2 \big)\,.$$

Proof. Assume without loss of generality that the first arm is optimal, i.e., $\mu_1 = \mu^*$ and $\Delta_1 = 0$. From (1), we have

$$\bar{R}_T = \sum_{k=1}^{K} \Delta_k \mathbb{E}\big[N_k(t)\big] .$$

Let $k \geq 2$ be a suboptimal arm. Then, the arm k is selected m times during the exploration phase, and $T - mK$ times during the exploitation if k is selected, which implies $\widehat{\mu}_k(m) \geq \widehat{\mu}_1(m)$. Therefore,

$$\mathbb{E}\big[N_k(t)\big] \leq m + (T - mK)\mathbb{P}\big(\widehat{\mu}_k(m) \geq \widehat{\mu}_1(m)\big)$$

Now, we can use Hoeffding's inequality to control the probability in the right-hand side. Indeed $\widehat{\mu}_k(m)$ and μ_1 are respectively the empirical averages of m i.i.d. random variables in $[0,1]$ of mean μ_k and $\mu_1 = \mu^*$. Therefore,

$$\begin{aligned} \mathbb{P}\big(\widehat{\mu}_k(m) - \widehat{\mu}_1(m) \geq 0\big) &= \mathbb{P}\big(\widehat{\mu}_k(m) - \widehat{\mu}_1(m) - \mu_k + \mu_1 \geq -\mu_k + \mu_1\big) \\ &= \mathbb{P}\big(\widehat{\mu}_k(m) - \widehat{\mu}_1(m) - \mu_k + \mu_1 \geq \Delta_k\big) \\ &= \mathbb{P}\big(m\widehat{\mu}_k(m) - m\widehat{\mu}_1(m) - m\mu_k + m\mu_1 \geq m\Delta_k\big) \\ &\leq \exp\big(-m\Delta_k^2\big) . \end{aligned}$$

This implies

$$\bar{R}_T \leq m \sum_{k=1}^{K} \Delta_k + (T - mK) \sum_{k=1}^{K} \Delta_k \exp\big(-m\Delta_k^2\big) .$$

□

The bound in Theorem 1 illustrates the trade-off between exploration and exploitation. If m is large, the exploration is too long and the first term $m\sum_{k=1}^{K}\Delta_k$ yields a large regret. On the other hand, for small m, there is a large probability to choose a suboptimal arm during the exploitation and the other term might lead to a large regret. The question is which value of m should we choose?
To have an idea, we will consider the case $K = 2$, in which case the bound is

$$\bar{R}_T \leq m\Delta_2 + T\Delta_2 \exp\big(-m\Delta_2^2\big) .$$

Corollary 1. *If $K = 2$ and $m = \max\big\{1, \lceil \log(T\Delta_2^2)/\Delta_2^2 \rceil\big\}$, then*

$$\bar{R}_T \leq \Delta_2 + \frac{1 + \log\big(T\Delta_2^2\big)}{\Delta_2} .$$

The above bound is of order $O((\log T)/\Delta_2)$. Such bounds are called distribution-dependent because they heavily depend on the distributions ν_k via Δ_k. If $\Delta_2 \to 0$, it explodes. However, we also have from (1) that $\bar{R}_T \leq \Delta_2 T$. Therefore, in the worst case for any value of Δ_2, Corollary 1 yields to the worst-case bound

$$\bar{R}_T \leq \min\left\{T\Delta_2, \Delta_2 + \frac{1 + \log\big(T\Delta_2^2\big)}{\Delta_2}\right\} \lesssim \sqrt{T \log T} . \tag{2}$$

The above bound is close to be optimal. Yet, the issue is that the parameter m depends on Δ_2 and T. If the dependence on T can be dealt with a doubling-trick it is harder to optimize it in Δ_2. Furthermore, when there are more than two arms, one might want to explore differently the arms. The upper-confidence-bound algorithm that we will see next does not have these issues.

Exercise 2.1. Show that it is possible to achieve the worst-case bound on the pseudo-regret of order $O(T^{2/3})$ by optimizing m independently of Δ (only with T).

Exercise 2.2. Generalize the results of Theorem 1 and 1 when the rewards are not-bounded but σ^2-sub-Gaussian, i.e., for all $\lambda > 0$

$$\mathbb{E}\left[\exp\big(\lambda(X_{k,t} - \mathbb{E}[X_{k,t}])\big)\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right) .$$

3 Upper-Confidence-Bound (UCB)

The UCB (Upper-Confidence-Bound) algorithm is a very popular bandit algorithm that has several advantages over ETC. It does not rely on an initial exploration phase but explores on the fly as rewards are observed. It explores and exploits sequentially throughout the experience. Unlike ETC, it does not require knowledge of gaps and behaves well when there are more than two arms.

To perform exploration and face uncertainty, the UCB algorithm is based on the *optimism principle*.

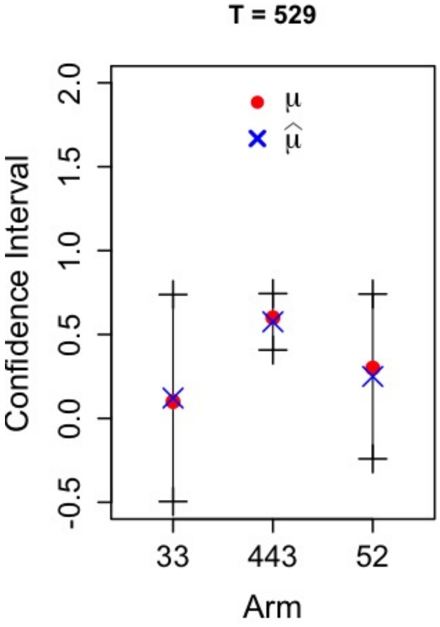
For each arm k, it builds a confidence interval on its expected reward based on past observation

$$I_k(t) = \big[LCB_k(k), UCB_k(t)\big]$$

where LCB is the Lower-Confidence-Bound and UCB is the Upper-Confidence-Bound. Then it is *optimistic* acting as if the best possible rewards are the real rewards and chooses the next arm accordingly

$$k_t \in \operatorname*{arg\,max}_{k \in \{1,\dots,K\}} \; UCB_k(t)\,.$$

In other words, it pulls the arm with the higher upper-confidence-bound. An example of how UCB works with three arms of means $\mu_1 = 0.1$, $\mu_2 = 0.6$ and $\mu_3 = 0.3$ is plotted in the Figure on the right. The best arm is pulled more often (see x-axis for number of times arms are selected) and his confidence interval is smaller.



The only question is how to design the upper-confidence-bounds. This is based on Hoeffding's inequality. Since the rewards are i.i.d. the distribution of $\widehat{\mu}_k(s)$ is equal to the distribution of

$$\frac{1}{s}\sum_{s'=1}^{s} X_{k,s'}\,,$$

with mean μ_k. Therefore, from Hoeffding's inequality, we have for all arms $k \in \{1,\dots,K\}$, for all $s \geq 1$ and all $\delta \in (0,1)$

$$\mathbb{P}\left\{\mu_k \geq \widehat{\mu}_k(s) + \sqrt{\frac{\log\frac{1}{\delta}}{2s}}\right\} \leq \delta\,. \tag{3}$$

where $\widehat{\mu}_k(t)$ is the empirical reward of arm k after pulling it t times. Therefore, it is reasonable to choose the upper-confidence bound

$$UCB_t(k) = \begin{cases} \infty & \text{if } N_k(t-1) = 0 \\ \widehat{\mu}_k(N_k(t-1)) + \sqrt{\frac{2\log t}{N_k(t-1)}} & \text{otherwise} \end{cases}$$

The UCB algorithm is described in Algorithm 2.

Theorem 2. *If the distributions ν_k have supports all included in $[0,1]$ then for all k such that $\Delta_k > 0$*

$$\mathbb{E}\big[N_k(T)\big] \leq \frac{8\log T}{\Delta_k^2} + 2\,.$$

```
**Initialization** For rounds $t = 1, \dots, K$ pull arm $k_t = t$
**For** $t = K + 1, \dots, T$, choose
    $k_t \in \mathop{\arg\max}_{k \in \{1,\dots,K\}} \left\{ \widehat{\mu}_k(N_k(t-1)) + \sqrt{\frac{2\log t}{N_k(t-1)}} \right\}$ ,
and get reward $X_{k_t,t}$.
```

Algorithm 2: Upper-Confidence-Bound (UCB)

In particular, this implies that the pseudo-regret of UCB is upper-bounded as

$$\bar{R}_T \le 2K + \sum_{k:\Delta_k>0} \frac{8\log T}{\Delta_k} \, .$$

Remark. Let us make some remarks about the about upper-bound on the pseudo-regret.

- UCB has a regret bound of order
$$\bar{R}_T \lesssim \frac{K \log T}{\Delta} \, ,$$
where $\Delta = \min_{i:\Delta_i>0} \Delta_i$. Once again, using that the regret incurred from pulling arm k cannot be larger than $T\Delta_k$, this distribution-dependent upper-bound can be transformed into a distribution-free bound of order $\bar{R}_T \lesssim \sqrt{TK\log T}$. We leave this proof as an exercise.
- This bound is close to optimal since the lower bound is of order $O(\sqrt{KT})$. There exists modification of UCB to get rid of the extra logarithmic term. For instance, the MOSS algorithm (Minimax Optimal Strategy in the Stochastic Case) achieves
$$\bar{R}_T \lesssim \min\left\{ \sqrt{TK}, \frac{K}{\Delta} \log \frac{T\Delta^2}{K} \right\} \, ,$$
however it depends on the smallest gap Δ only and not on all gaps Δ_i.
- The assumption that the rewards are independent between arms can be relaxed.
- The assumption that the rewards are in $[0,1]$ can be relaxed to a sub-Gaussian assumption.
- While a bound on the pseudo-regret is interesting, one would actually want a bound with high probability on
$$\widehat{R}_T \overset{\text{def}}{=} T\mu^* - \sum_{t=1}^{T} \mu_{k_t,t} \, .$$
Using Hoeffding's inequality to control $\widehat{R}_T$ with $\bar{R}_T = \mathbb{E}[\widehat{R}_T]$ would yield an additional term of order $\sqrt{T}$ due to fluctuations which would dominate $O(K\log T/\Delta)$. Obtaining a bound of order $O(K\log T/\Delta)$ on $\widehat{R}_T$ is a challenging problem and not achieved by UCB. Some strategies using the knowledge of T can satisfy it.

Proof. Without loss of generality let us assume that the first arm is optimal, i.e., $\mu_1 = \mu^*$ and $\Delta_1 = 0$. We show below that if $k_t = k$, then at least one of the following three inequalities must be satisfied

$$\mu^* > \widehat{\mu}_1\big(N_1(t-1)\big) + \sqrt{\frac{2\log t}{N_1(t-1)}} \qquad \leftarrow \mu^* \text{ larger than UCB} \tag{i}$$

$$\mu_k < \widehat{\mu}_k\big(N_k(t-1)\big) - \sqrt{\frac{2\log t}{N_k(t-1)}} \qquad \leftarrow \mu_k \text{ smaller than LCB} \tag{ii}$$

$$N_k(t-1) \le \frac{8\log t}{\Delta_k^2} \qquad \leftarrow k \text{ not played enough yet} \tag{iii}$$

Indeed, otherwise assume that the three inequalities are all false than

$$\begin{aligned}
\widehat{\mu}_1(N_1(t-1)) + \sqrt{\frac{2\log t}{N_1(t-1)}} &\geq \mu^* && \leftarrow \quad \text{not (i)} \\
&\geq \mu_k + \Delta_k && \leftarrow \quad \text{Def of } \Delta_k \\
&> \mu_k + 2\sqrt{\frac{2\log t}{N_k(t-1)}} && \leftarrow \quad \text{not (iii)} \\
&\geq \widehat{\mu}_k(N_k(t-1)) + \sqrt{\frac{2\log t}{N_k(t-1)}} && \leftarrow \quad \text{not (ii)}\,.
\end{aligned}$$

This contradicts the fact that $k_t = k$ (see Algorithm 2). Therefore, denoting $u = \lfloor \frac{8\log T}{\Delta_k^2} \rfloor$, we have

$$\begin{aligned}
\mathbb{E}\big[N_k(T)\big] = \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}_{k_t = k}\right] &= u + \sum_{t=u+1}^{T} \mathbb{P}\Big\{ k_t = k \text{ and (iii) is false } \Big\} \\
&= u + \sum_{t=u+1}^{T} \Big(\mathbb{P}\big\{\text{(i) or (ii)}\big\} \Big) \\
&\leq u + \sum_{t=u+1}^{T} \Big(\mathbb{P}\big\{\text{(i)}\big\} + \mathbb{P}\big\{\text{(ii)}\big\} \Big)\,. \qquad (4)
\end{aligned}$$

Therefore, it suffices to control the probabilities of (i) and (ii), which we do now. At round $t \geq 1$,

$$\begin{aligned}
\mathbb{P}\big\{\text{(i)}\big\} &\leq \mathbb{P}\left\{ \exists s \in \{1,\dots,t-1\}, \text{ such that } \ \mu^* > \widehat{\mu}_1(s) + \sqrt{\frac{2\log t}{s}} \right\} \\
&\leq \sum_{s=1}^{t} \mathbb{P}\left\{ \mu^* > \widehat{\mu}_1(s) + \sqrt{\frac{\log(1/t^{-4})}{2s}} \right\} \\
&\overset{(3)}{\leq} \sum_{s=1}^{t} t^{-4} = t^{-3}\,.
\end{aligned}$$

By symmetry, the same applies for $\mathbb{P}\big\{\text{(ii)}\big\} \leq t^{-3}$. Combining into (4), it concludes the proof of the first inequality

$$\mathbb{E}\big[N_k(T)\big] \leq \frac{8\log T}{\Delta_k^2} + 2\sum_{t=u+1}^{T} t^{-3} \leq \frac{8\log T}{\Delta_k^2} + 2\,.$$

The upper-bound on the pseudo-regret follows from (1). □

4 Other algorithms

Other algorithms exist in the literature. The best known are ε-greedy and Thompson sampling.

4.1 ε-greedy

The idea of ε-greedy is very simple: first choose a parameter $\varepsilon \in (0,1)$, then at each round, select the arm with the highest empirical mean with probability ε (i.e., be greedy), and explore by playing a random arm with probability ε. It works quite well in practice and is used in many application because of its simple implementation (in particular in reinforcement learning). Choosing $\varepsilon \approx K/(\Delta^2 T)$ yields to an upper-bound of order $K\log T/\Delta^2$. However it requires the knowledge of Δ.

4.2 Thompson Sampling

Thomson sampling was the first algorithm proposed for bandits by Thomson in 1933. It assumes a uniform prior over the expected rewards $\mu_i \in (0,1)$, then at each round $t \geq 1$, for each arm $\pi_{k,t}$, it

- computes $\widehat{\nu}_{k,t}$ the posterior distribution of the rewards of arm k given the rewards observed so far;
- samples $\theta_{k,t} \sim \widehat{\nu}_{k,t}$ independently;
- selects $k_t \in \arg\max_{k \in \{1,\ldots,K\}} \theta_{k,t}$.

Thomson sampling has a similar upper-bound of order $O(K \log T/\Delta)$ than the one achieved by UCB. An advantage over UCB is the possibility of incorporating easily prior knowledge on the arms.

References

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint*, 2019.