

Sequential learning – Lesson 3

Stochastic Bandits

Rémy Degenne

February 3, 2023

Centre Inria de l'Université de Lille

The stochastic bandit problem

Consequences of the stochasticity

Exploration and exploitation

Optimism in face of uncertainty: UCB

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

- the player observes a context $x_t \in \mathcal{X}$ (optional step)
- the player chooses an action $\theta_t \in \Theta$ (compact decision/parameter set);
- the environment chooses a loss function $\ell_t : \Theta \rightarrow [0, 1]$;
- the player suffers loss $\ell_t(\theta_t)$ and observes
 - the losses of every actions: $\ell_t(\theta)$ for all $\theta \in \Theta$ \rightarrow full-information feedback
 - the loss of the chosen action only: $\ell_t(\theta_t)$ \rightarrow bandit feedback.

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

$$\hat{L}_T \stackrel{\text{def}}{=} \sum_{t=1}^T \ell_t(\theta_t).$$

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

- the player observes a context $x_t \in \mathcal{X}$ (optional step)
- the player chooses an arm $k_t \in \Theta$ (compact decision/parameter set, most often $\{1, \dots, K\}$);
- the player observes
 - the **rewards** of every arm: $X_t^k \sim \nu_k$ for all $k \in \Theta$ \rightarrow full-information feedback
 - the **reward** of the chosen arm only: $X_t^{k_t} \sim \nu_{k_t}$ \rightarrow bandit feedback.

The goal of the player is to maximize their cumulative reward.

Regret?

We could use the definition of the regret from adversarial bandits:

Definition (Regret, attempt 1)

$$R_T = \max_k \sum_{t=1}^T X_t^k - \sum_{t=1}^T X_t^{k_t}.$$

Let's see why we don't use that definition.

Notations and assumptions:

- The arm set is $[K] = \{1, \dots, K\}$.
- $\mu^k = \mathbb{E}_{X \sim \nu_k}[X]$, assumed finite for all arms k .
- $\mu^* = \max_{k \in [K]} \mu^k$.

The first notion of regret is inadequate

$$R_T = \max_k \sum_{t=1}^T X_t^k - \sum_{t=1}^T X_t^{k_t}.$$

ν_k Bernoulli(1/2) for all $k \in [K]$. $\mu^k = 1/2$ for all k .

All arms are the same \rightarrow there is no bad choice and **no bad algorithm**.

But:

$$\begin{aligned}\mathbb{E}R_T &= \mathbb{E}[\max_{k \in [K]} \sum_{t=1}^T X_t^k] - T/2 \\ &= \mathbb{E}[\max_{k \in [K]} \sum_{t=1}^T (X_t^k - 1/2)] \\ &\approx \sqrt{T \log K}\end{aligned}$$

Regret definition

We want a regret notion that does not blow up with stochastic fluctuations.

Definition ((Pseudo)-Regret)

The regret is defined as

$$R_T = \max_k \sum_{t=1}^T \mu^k - \sum_{t=1}^T \mu^{k_t} = T\mu^* - \sum_{t=1}^T \mu^{k_t}.$$

Recall that $\mu^k = \mathbb{E}_{X \sim \nu_k}[X]$.

Most often, we bound the **expected regret** $\mathbb{E}[R_T]$.

Note that the expectation here is over the random rewards and the randomness of the algorithm, if there is any.

Regret decomposition

Suppose that the set of arms is finite: $[K]$.

Define the **gap** of arm $k \in [K]$ by $\Delta_k = \mu^* - \mu^k$.

$$R_T = T\mu^* - \sum_{t=1}^T \mu^{k_t} = \sum_{t=1}^T (\mu^* - \mu^{k_t}) = \sum_{t=1}^T \Delta_{k_t} = \sum_{k=1}^K N_T^k \Delta_k ,$$

where $N_T^k = \sum_{t=1}^T \mathbb{I}\{k_t = k\}$ is the number of pulls of arm k up to time T .

Bounding the regret \Leftrightarrow bounding the number of pulls of bad arms

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

- the player observes a context $x_t \in \mathcal{X}$ (optional step)
- the player chooses an arm $k_t \in \Theta$ (compact decision/parameter set, most often $\{1, \dots, K\}$);
- the player observes the reward of the chosen arm only: $X_t^{k_t} \sim \nu_{k_t}$ (independent of other rewards).

The goal of the player is to minimize their expected regret: $\mathbb{E}[R_T] = \sum_{k=1}^K \mathbb{E}[N_T^k] \Delta_k$.

The setting given before is quite informal.

We described:

- Conditional distributions of the rewards (Markov kernels)
- Conditional distributions of the pulls

Is there a probability space over $(k_1, X_1^{k_1}, k_2, X_2^{k_2}, \dots)$ compatible with those specifications?

See Lattimore and Szepesvári, “Bandit algorithms”, 2019, section 4.6, for a formal construction using the Ionescu-Tulcea theorem.

Variants and extensions

Setting variants:

- **Contextual bandit:** $X_t^{k_t} \sim \nu_{k_t}(x_t)$, for a known context x_t
- **Linear bandit:** $\nu_{k_t} = \mathcal{N}(\theta^\top x_{k_t}, 1)$
- **Structured bandit:** the algorithm knows constraints on $(\mu^k)_{k \in [K]}$, e.g. Lipschitz, linear, monotone...

Goal variants: instead of minimizing the regret, we want to

- **Minimize the simple regret:** return an arm at time T , and minimize its expected gap.
- **Identify the best arm:** return an arm at time T , and minimize the probability that its not one of the best ones.

Relaxed assumptions: rewards not independent, distributions changing with time, etc.

The stochastic bandit problem

Consequences of the stochasticity

Exploration and exploitation

Optimism in face of uncertainty: UCB

Convergence to the mean

Main idea: we can estimate the mean of the arms with the empirical mean.

Let $(X_s)_{s \in \mathbb{N}}$ be iid random variables with $\mathbb{E}[|X_1|] < \infty$ and expected value $\mathbb{E}[X_1] = \mu$.

Let $\bar{X}_t = \sum_{s=1}^t X_s$.

Theorem 1 (Strong law of large numbers)

$\bar{X}_t \xrightarrow{a.s.} \mu$, that is $\mathbb{P}(\bar{X}_t \rightarrow \mu) = 1$.

Theorem 2 (Central limit theorem)

If $\mathbb{V}[X] = \sigma^2 < \infty$, then $\sqrt{t}(\bar{X}_t - \mu) \xrightarrow{d} \mathcal{N}(\mu, \sigma^2)$.

Problem: those are asymptotic results.

Main question: if I have 15 samples of arm k , how reliable is my estimate for μ^k ?

Concentration inequalities

Our main tools are **concentration inequalities**: bounds on the probability that the empirical mean (or another statistic) is far from its expected value.

Theorem 3 (Hoeffding's inequality)

If X_1, \dots, X_t are independent random variables almost surely in $[a, b]$ then for all $\delta \in (0, 1)$ we have

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

Equivalently, for all $\varepsilon \geq 0$,

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

Proof under a sub-Gaussian assumption. Bounded implies sub-Gaussian.

Assumption: for all s , X_s is σ^2 -sub-Gaussian, which means that for all $\lambda \in \mathbb{R}$,
$$\mathbb{E}[e^{\lambda(X_s - \mu_s)}] \leq e^{\frac{1}{2}\sigma^2\lambda^2}.$$

Warning: random number of samples

In the analysis of bandit algorithms, we want to bound $\hat{\mu}_t^k - \mu^k$, where

$$\hat{\mu}_t^k = \frac{1}{N_t^k} \sum_{s=1}^t X_s^{k_s} \mathbb{I}\{k_s = k\}.$$

k_s is a random variable that depends on all previous rewards.

Issue: $\hat{\mu}_t^k$ is a sum of a **random number** of random variables.

- $\hat{\mu}_t^k$ is **not** unbiased.
- $\hat{\mu}_t^k$ is **not** a sum of a fixed number of independent random variables.
- **Hoeffding's inequality does not apply.**

How to avoid the difficulty: union bounds, or martingale arguments (see proofs later in the course).

The stochastic bandit problem

Consequences of the stochasticity

Exploration and exploitation

Optimism in face of uncertainty: UCB

Follow the leader

Goal: minimize $\mathbb{E}[R_T] = T\mu^* - \sum_{t=1}^T \mu^{k_t}$.

Since the empirical mean of an arm concentrate around its expected value, can we simply pull the arm with highest empirical mean?

Definition (Follow-The-Leader)

The FTL algorithm pulls arm $k_t = \arg \max_{k \in [K]} \hat{\mu}_{t-1}^k$.

Full information: yes, FTL is optimal.

Bandit: answer is **no**, FTL does not work. It has **linear expected regret** in most settings.

Explore then commit

Need to not only **exploit**, but also **explore**.

Explore-Then-Commit

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

$$\hat{\mu}_{mK}^k = \frac{1}{m} \sum_{s=1}^{Km} \mathbb{I}\{k_s = k\} X_s^k.$$

2. Exploitation: keep playing the best arm $\arg \max_k \hat{\mu}_{mK}^k$ for the remaining rounds $t = mK + 1, \dots, T$.

Theorem 4 (Thm 6.1, Lattimore and Szepesvári, “Bandit algorithms”, 2019)

If all distributions are bounded in $[0, 1]$ and $1 \leq m \leq T/K$ then ETC has expected regret

$$\mathbb{E}[R_T] \leq m \sum_{k=1}^K \Delta_k + (T - mK) \sum_{k=1}^K \Delta_k \exp(-m\Delta_k^2).$$

- m too large \Rightarrow too much exploration, linear regret.
- m too small \Rightarrow too little exploration, linear regret.
- What m should we choose?

Finding the right trade-off

Two arms bandit: arm 1 is the best arm, arm 2 has gap Δ .

ETC verifies

$$\mathbb{E}[R_T] \leq m\Delta + (T - 2m)\Delta e^{-m\Delta^2}.$$

Theorem 5

If $K = 2$ and $m = \max\{1, \left\lceil \log \frac{\log(T\Delta^2)}{\Delta^2} \right\rceil\}$, then

$$\mathbb{E}[R_T] \leq \Delta + \frac{1 + \log(T\Delta^2)}{\Delta}.$$

This is a **distribution dependent** bound, meaning that it depends on the gap.

Issue with those bounds: meaningless if Δ is small.

Worst case bound

ETC verifies

$$\mathbb{E}[R_T] \leq m\Delta + (T - 2m)\Delta e^{-m\Delta^2}.$$

Theorem 6

If $K = 2$ and $m = \max\{1, \lceil \log \frac{\log(T\Delta^2)}{\Delta^2} \rceil\}$, then

$$\mathbb{E}[R_T] \leq \min \left\{ \Delta + \frac{1 + \log(T\Delta^2)}{\Delta}, T\Delta \right\} \lesssim \sqrt{T \log T}.$$

This is close to optimal: we can prove a lower bound of order \sqrt{T} .

Problems:

- m depends on Δ , which is unknown.
- What can we do for $K > 2$?

The stochastic bandit problem

Consequences of the stochasticity

Exploration and exploitation

Optimism in face of uncertainty: UCB

Being optimistic

The UCB (Upper-Confidence-Bound) algorithm is a very popular bandit algorithm that has several advantages over ETC:

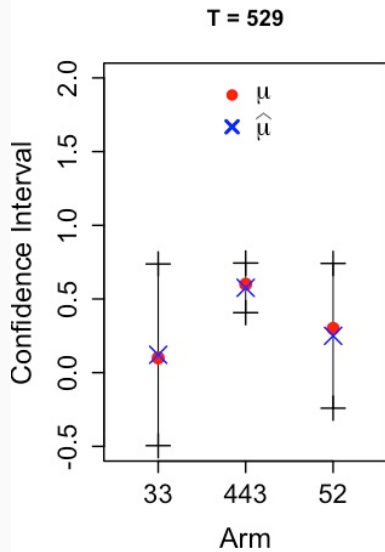
- It does not rely on an initial exploration phase but explores on the fly as rewards are observed.
- Unlike ETC, it does not require knowledge of gaps and behaves well when there are more than two arms.

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

$$I_t^k = [L_t^k, U_t^k] .$$

It is **optimistic**, acting as if the best possible rewards are the real rewards:

$$k_t \in \arg \max_{k \in \{1, \dots, K\}} U_t^k .$$



Confidence intervals

How to design the upper confidence bounds?

→ concentration inequalities. Here **Hoeffding's inequality**.

Theorem 7 (Hoeffding's inequality)

If X_1, \dots, X_t are independent random variables almost surely in $[a, b]$ with same mean μ then for all $\delta \in (0, 1)$ we have

$$\mathbb{P} \left(\frac{1}{t} \sum_{s=1}^t X_s - \mu \geq \sqrt{\frac{(b-a)^2}{2t} \log \frac{1}{\delta}} \right) \leq \delta .$$

Careful: UCB is adaptive, hence $\hat{\mu}_t$ is not exactly a sum of independent random variables. But we will make it work.

For rewards in $[0, 1]$: $U_t^k = \hat{\mu}_{t-1}^k + \sqrt{\frac{2 \log t}{N_{t-1}^k}}$

Initialization For rounds $t = 1, \dots, K$ pull arm $k_t = t$.

For $t = K + 1, \dots, T$, choose

$$k_t \in \arg \max_{k \in [K]} \left\{ \hat{\mu}_{t-1}^k + \sqrt{\frac{2 \log t}{N_{t-1}^k}} \right\},$$

and get reward $X_t^{k_t}$.

Theorem 8

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

$$\mathbb{E}[N_T^k] \leq \frac{8 \log T}{\Delta_k^2} + 2.$$

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

$$\mathbb{E}[R_T] \leq \sum_{k: \Delta_k > 0} \frac{8 \log T}{\Delta_k} + 2 \sum_{k=1}^K \Delta_k.$$

Remarks :

- we can also prove $\mathbb{E}[R_T] \lesssim \sqrt{KT \log(T)}$. Close to the optimal $O(\sqrt{KT})$.
- Deals with multiple gaps, without any knowledge of the gaps, unlike ETC.
- Bounded can be replaced by sub-Gaussian.

Proof start

Idea: if the means belong to the confidence intervals and the arms are pulled enough, the algorithm cannot pull a suboptimal arm.

We prove that if $k_t = k \neq *$, then one of these inequalities must be false:

$$\mu^* \leq \hat{\mu}_{t-1}^* + \sqrt{\frac{2 \log t}{N_{t-1}^*}} \quad \leftarrow \mu^* \text{ smaller than UCB} \quad (\text{i})$$

$$\mu^k \geq \hat{\mu}_{t-1}^k - \sqrt{\frac{2 \log t}{N_{t-1}^k}} \quad \leftarrow \mu_k \text{ larger than LCB} \quad (\text{ii})$$

$$N_{t-1}^k \geq \frac{8 \log t}{\Delta_k^2} \quad \leftarrow k \text{ played enough} \quad (\text{iii})$$

Proof 2

$$\mu^* \leq \hat{\mu}_{t-1}^* + \sqrt{\frac{2 \log t}{N_{t-1}^*}} \quad \text{and} \quad \mu^k \geq \hat{\mu}_{t-1}^k - \sqrt{\frac{2 \log t}{N_{t-1}^k}} \quad \text{and} \quad N_{t-1}^k \geq \frac{8 \log t}{\Delta_k^2}$$

Prove that if k is pulled at t , then there is a contradiction.

Proof 3: decomposition wrt events

One of these is false:

$$\mu^* \leq \hat{\mu}_{t-1}^* + \sqrt{\frac{2 \log t}{N_{t-1}^*}} \quad ; \quad \mu^k \geq \hat{\mu}_{t-1}^k - \sqrt{\frac{2 \log t}{N_{t-1}^k}} \quad ; \quad N_{t-1}^k \geq \frac{8 \log t}{\Delta_k^2}$$

Then: $\mathbb{E}[N_T^k] \leq u + \sum_{t=u+1}^T \left(\mathbb{P}\{(i) \text{ is false}\} + \mathbb{P}\{(ii) \text{ is false}\} \right) ..$

Proof 4: probability of the concentration event

We show: $\mathbb{P}(\mu^k < \hat{\mu}_{t-1}^k - \sqrt{\frac{2 \log t}{N_{t-1}^k}}) \leq t^{-3}.$

For $u = \frac{8 \log T}{\Delta_k^2}$, $\mathbb{E}[N_T^k] \leq u + \sum_{t=u+1}^T \left(\mathbb{P}\{\mu^* > \hat{\mu}_{t-1}^* + \sqrt{\frac{2 \log t}{N_{t-1}^*}}\} + \mathbb{P}\{\mu^k < \hat{\mu}_{t-1}^k - \sqrt{\frac{2 \log t}{N_{t-1}^k}}\} \right)$.

Each of these probabilities is smaller than t^{-3} .

$$\mathbb{E}[N_T^k] \leq \frac{8 \log T}{\Delta_k^2} + 2 \sum_{t=u+1}^T \frac{1}{t^3} \leq \frac{8 \log T}{\Delta_k^2} + 2.$$

The bound of the regret then comes from $\mathbb{E}[R_T] = \sum_k \mathbb{E}[N_T^k] \Delta_k$.

ε -greedy

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 .

Works quite well in practice and is used in many applications 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 .

Other Algorithms: Thompson Sampling

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 $t \geq 1$, it





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

Thomson sampling has a similar upper-bound of order $O(K \log T / \Delta)$ than the one achieved by UCB. Somewhat different proof techniques.

An advantage over UCB is the possibility of incorporating easily prior knowledge on the arms.

UCB proved easier to adapt to structured bandits (it can be hard to sample a posterior conditioned on structural information).

Thank you!

-  Cesa-Bianchi, Nicolo and Gábor Lugosi. Prediction, learning, and games. Cambridge university press, 2006.
-  Hazan, Elad et al. “Introduction to online convex optimization”. In: Foundations and Trends® in Optimization 2.3-4 (2016), pp. 157–325.
-  Lattimore, Tor and Csaba Szepesvári. “Bandit algorithms”. In: preprint (2019).
-  Shalev-Shwartz, Shai et al. “Online learning and online convex optimization”. In: Foundations and Trends® in Machine Learning 4.2 (2012), pp. 107–194.

