Online Learning Applications

Part 2: Stochastic MABs

Stochastic MABs

At each time $t = 1, \ldots, T$:

1. The reward $r_t(a)$ of an **arm** a is sampled from a distribution $\mathcal{D}_a$ supported on $[0, 1]$

Stochastic MABs

At each time $t = 1, \ldots, T$:

1. The reward $r_t(a)$ of an **arm** a is sampled from a distribution $\mathcal{D}_a$ supported on $[0, 1]$ (**alternatively we can assume subgaussian noise**)

Stochastic MABs

At each time $t = 1, \ldots, T$:

1. The reward $r_t(a)$ of an **arm** a is sampled from a distribution $\mathcal{D}_a$ supported on $[0, 1]$
2. The learner chooses an arm $a_t \in A$
3. The learner receives a reward $r_t(a_t)$
4. The learner observes only the reward $r_t(a_t)$ of arm a_t

Stochastic MABs

At each time $t = 1, \ldots, T$:

1. The reward $r_t(a)$ of an **arm** a is sampled from a distribution $\mathcal{D}_a$ supported on $[0, 1]$
2. The learner chooses an arm $a_t \in A$
3. The learner receives a reward $r_t(a_t)$
4. The learner observes only the reward $r_t(a_t)$ of arm a_t

Recall that $r_t(a) = 1 - \ell_t(a)$. We use reward for the stochastic setting and loss for adversarial one. This is just to be consistent with the literature and textbooks.

Stochastic MABs

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

1. The reward $r_t(a)$ of an **arm** a is sampled from a distribution $\mathcal{D}_a$ supported on $[0, 1]$
2. The learner chooses an arm $a_t \in A$
3. The learner receives a reward $r_t(a_t)$
4. The learner observes only the reward $r_t(a_t)$ of arm a_t

Recall that $r_t(a) = 1 - \ell_t(a)$. We use reward for the stochastic setting and loss for adversarial one. This is just to be consistent with the literature and textbooks.

Goal

Design an algorithm that achieves **sublinear pseudo-regret** ($\lim_{T\to\infty} \frac{\mathcal{R}_T}{T} = 0$).

Some definitions

We need to define some quantities:

- $\mu(a) = \mathbb{E}_{r \sim \mathcal{D}_a} r(a)$
- $a^* \in \arg\max_a \mu(a)$ is an **optimal** arm
- For any arm a, we define the **sub-optimality gap** as:

$$\Delta_a = \mu(a^*) - \mu(a)$$

Pseudo-regret

The pseudo-regret of an algorithm is:

$$\mathcal{R}_T = T\mu(a^*) - \mathbb{E}\left[\sum_{t \in [T]} \mu(a_t)\right],$$

where the expectation is on the randomness of the algorithm.

Regret decomposition

$N_a(t)$: the number of time we pick arm a in the first t rounds:

$$N_a(t) := \sum_{j=1}^{t} \mathbb{I}[a_j = a].$$

Regret decomposition lemma

Given a stochastic expert problem,

$$\mathcal{R}_T = \sum_{a \in A} \Delta_a \mathbb{E}[N_a(T)].$$

Regret decomposition

Proof.

$$
\begin{aligned}
\mathcal{R}_T &:= T\mu(a^*) - \mathbb{E}\left[\sum_{t\in[T]} \mu(a_t)\right] \\
&= T\mu(a^*) - \sum_{t\in[T]} \mathbb{E}\left[\mu(a_t)\right] \\
&= \sum_{t\in[T]} \left(\mu(a^*) - \mathbb{E}\left[\mu(a_t)\right]\right) \\
&= \sum_{a\in A} \mathbb{E}\left[N_a(T)\right]\left(\mu(a^*) - \mu(a)\right) \\
&= \sum_{a\in A} \Delta_a \mathbb{E}[N_a(T)].
\end{aligned}
$$

$\square$

First idea: greedy

Greedy

At each round $t = 1, \ldots, T$:

1. Estimate the average reward as $\mu_t(a) = \frac{1}{N_{t-1}(a)} \sum_{t'=1}^{t-1} r_{t'}(a)\mathbb{I}[a_{t'} = a]$
2. (If an arm isn't been played then we set $\mu_t(a) = \infty$)
3. Select the arm with highest estimated reward $\mu_t(a)$

Greedy suffers linear regret

Theorem

The greedy algorithm suffers regret $\Omega(T)$.

Proof sketch.

- Two arms a_1, a_2
- The reward of arm a_1 is $r_t(a_1) = 0$ with probability $1/2$ and $r_t(a_1) = 1$ with probability $1/2$.
- The reward of arm a_2 is always $r_t(a_2) = 1/4$
- With probability $1/2$, when we play the arm a_1 for the first time the reward is 0 and the empirical mean is $\mu_t(a_1) = 0$
- The empirical mean of arm a_2 is always $\mu_t(a_2) = 1/4$
- The algorithm will never play the optimal arm a_1 again!
- We suffer regret $1/4$ at each round

$\square$

A better idea: Explore-Then-Commit (ETC)

- We need to **explore** more!
- We can explore uniformly for $KT_0 < T$ rounds and then commit to the arm with the best empirical mean

Explore-Then-Commit

Given $T_0 \in \{1, \ldots, T/K\}$

1. Play each arm $a \in A$ for T_0 times.
2. At round $\hat{t} = KT_0 + 1$, compute the arm with the best empirical mean

$$\hat{a} = \arg\max_{a} \mu_{\hat{t}}(a)$$

3. Play arm $\hat{a}$ for all $t \geq \hat{t}$ (*i.e.*, until the end)

Analysis of ETC

Theorem

Explore-then-commit with $T_0 = (T/K)^{2/3} \log(T)^{1/3}$ guarantees

$$\mathcal{R}_T = O(T^{2/3}(K \log(T))^{1/3})$$

For simplicity, we provide an analysis for two arms a_1 and a_2.

- W.l.o.g., we assume that a_1 is the optimal arm
- In the exploration rounds, the alg. incurs expected pseudo-regret Δ_{a_2} every time arm a_2 is played
- In the exploration rounds, the expected pseudo-regret is $T_0 \Delta_{a_2} \leq T_0$
- In the commit rounds, the alg. incurs expected pseudo-regret $\Delta_{a_2}(T - KT_0)$ with probability $\mathbb{P}[\mu_{\hat{t}}(a_2) \geq \mu_{\hat{t}}(a_1)]$

Analysis of ETC

Theorem

Explore-then-commit with $T_0 = (T/K)^{2/3} \log(T)^{1/3}$ guarantees

$$\mathcal{R}_T = O(T^{2/3}(K \log(T))^{1/3})$$

For simplicity, we provide an analysis for two arms a_1 and a_2.

- W.l.o.g., we assume that a_1 is the optimal arm
- In the exploration rounds, the alg. incurs expected pseudo-regret Δ_{a_2} every time arm a_2 is played
- In the exploration rounds, the expected pseudo-regret is $T_0\Delta_{a_2} \leq T_0$
- In the commit rounds, the alg. incurs expected pseudo-regret $\Delta_{a_2}(T - KT_0)$ with probability $\mathbb{P}[\mu_{\hat{t}}(a_2) \geq \mu_{\hat{t}}(a_1)]$
- **How to upper bound this probability?**

Concentration inequalities

- $\mu_{\hat{t}}(a)$ is the average of T_0 i.i.d random variables (i.e., sampled from the same distribution $\mathcal{D}_a$)
- We want $\mu_{\hat{t}}(a)$ to be "close" to the true mean $\mu(a)$
- The Hoeffding inequality bounds the **probability that the empirical mean is far from the actual mean**

Hoeffding Inequality

Let $r_1, \ldots, r_n \in [0, 1]$ be i.i.d. random variables with mean μ. Then, for each $\epsilon > 0$

$$\mathbb{P}\left[\left|\frac{\sum_{i\in[n]} r_i}{n} - \mu\right| \geq \epsilon\right] \leq 2e^{-2\epsilon^2 n}$$

Concentration inequalities

- $\mu_{\hat{t}}(a)$ is the average of T_0 i.i.d random variables (i.e., sampled from the same distribution $\mathcal{D}_a$)
- We want $\mu_{\hat{t}}(a)$ to be "close" to the true mean $\mu(a)$
- The Hoeffding inequality bounds the **probability that the empirical mean is far from the actual mean**

Hoeffding Inequality

Let $r_1, \dots, r_n \in [0,1]$ be i.i.d. random variables with mean μ. Then, for each $\epsilon > 0$

$$\mathbb{P}\left[\left|\frac{\sum_{i\in[n]} r_i}{n} - \mu\right| \geq \epsilon\right] \leq 2e^{-2\epsilon^2 n}$$

- The probability decreases exponentially in ϵ^2 $\rightarrow$ better approximation implies larger error probability
- The probability decreases exponentially in the number of samples n

Analysis of ETC

- Hoeffding Inequality on $\mu(a_1)$: with probability $1 - 2e^{-2\epsilon^2 T_0}$

$$\mu_{\hat{t}}(a_1) \geq \mu(a_1) - \epsilon$$

- Hoeffding Inequality on $\mu(a_2)$: with probability $1 - 2e^{-2\epsilon^2 T_0}$

$$\mu_{\hat{t}}(a_2) \leq \mu(a_2) + \epsilon$$

- By an union bound, both inequalities hold simultaneously with probability at least $1 - 4e^{-2\epsilon^2 T_0}$

Assume that both inequalities hold and we commit to the sub-optimal arm a_2.

$$\mu(a_2) \geq \mu_{\hat{t}}(a_2) - \epsilon \geq \mu_{\hat{t}}(a_1) - \epsilon \geq \mu(a_1) - 2\epsilon$$

Analysis of ETC

The total regret in the commit phase is:

- With probability at least $1 - 4e^{-2\epsilon^2 T_0}$, at most $2\epsilon(T - KT_0)$
- With probability at most $4e^{-2\epsilon^2 T_0}$, at most $T - KT_0$

The total regret is at most:

$$\mathcal{R}_T \leq T_0 + (1 - 4e^{-2\epsilon^2 T_0})2\epsilon(T - KT_0) + 4e^{-2\epsilon^2 T_0}(T - KT_0)$$

Setting:

- $T_0 = T^{2/3}(\log(T))^{1/3}$
- $\epsilon = \sqrt{\frac{\log T}{T_0}}$

$$\mathcal{R}_T \leq O\left(T^{2/3}(\log(T)^{1/3})\right)$$

Lower bounds

Question

How can understand if the $\widetilde{O}(T^{2/3})$ bound we just derived is optimal, or whether we may be able to design a bandit algorithm with better regret guarantees?

Goal: lower bounds on regret which apply to **all** bandits algorithms at once.

Lower bounds

Theorem

Any algorithm suffers pseudo-regret at least

$$\mathcal{R}_T \geq \Omega\left(\sqrt{KT}\right).$$

Worst-case lower bound: an algorithm may have better regret on specific problems

Theorem

Any algorithm suffers pseudo-regret at least

$$\mathcal{R}_T \geq \Omega\left(\sum_{a:\Delta_a>0} \frac{1}{\Delta_a} \log(T)\right).$$

UCB1

UCB1

- To achieve optimal regret bounds we need to **explore** and **exploit** at the same time
- We can be greedy but use **optimism** to incentivize exploration
- The general idea of UCB1 is to:
  - Define an Upper Confidence Bound (UCB) on the expected mean of each arm
  - At each round, play the arm with the higher UCB $\rightarrow$ **here the algorithm is optimistic about the mean incentivizing exploration)**
  - The UCB of the played arm is updated

Upper Confidence Bound

For each arm $a \in A$, we build a confidence interval around its empirical mean.

At each time $t \in [T]$, we define an UCB of the arm average reward $\mu_t(a)$:

$$UCB_t(a) = \underbrace{\mu_t(a)}_{exploitation\ \ term} + \underbrace{\sqrt{\frac{2\log(T)}{N_{t-1}(a)}}}_{exploration\ \ term}$$

- The term $\mu_t(a)$ incentivizes to play arms with large empirical mean
- The term $\sqrt{\frac{2\log(T)}{N_{t-1}(a)}}$ incentivizes to play arms with low $N_{t-1}(a) \rightarrow$ played a small number of times

Upper Confidence Bound

For each arm $a \in A$, we build a confidence interval around its empirical mean.

At each time $t \in [T]$, we define an UCB of the arm average reward $\mu_t(a)$:

$$UCB_t(a) = \underbrace{\mu_t(a)}_{exploitation\ \ term} + \underbrace{\sqrt{\frac{2\log(T)}{N_{t-1}(a)}}}_{exploration\ \ term}$$

- The term $\mu_t(a)$ incentivizes to play arms with large empirical mean
- The term $\sqrt{\frac{2\log(T)}{N_{t-1}(a)}}$ incentivizes to play arms with low $N_{t-1}(a) \rightarrow$ played a small number of times

⚠ We can also take the smaller exploration term $\sqrt{\frac{2\log(t)}{N_{t-1}(a)}}$.

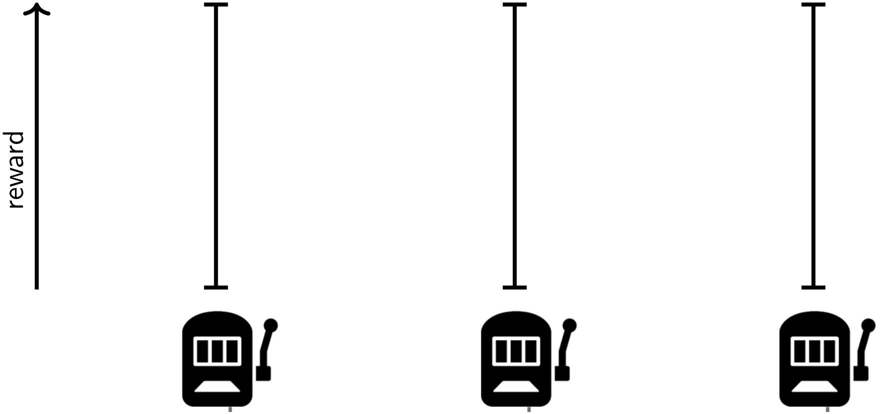
- More complex theoretical analysis and same regret bound
- It might have better empirical performances

UCB1

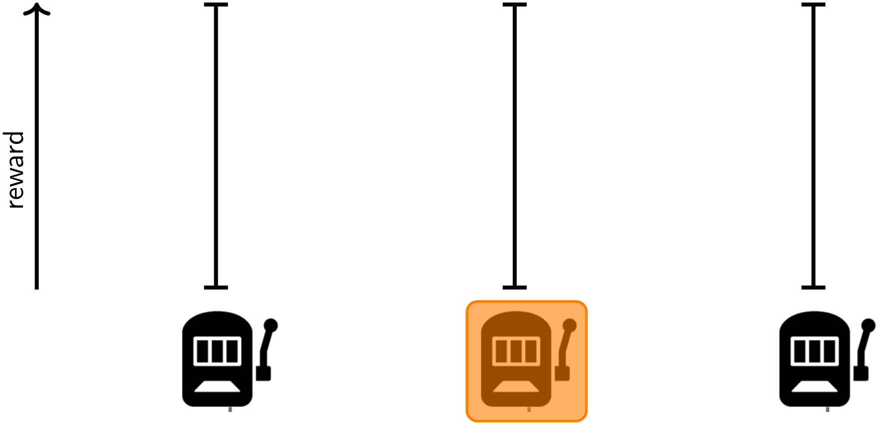
Algorithm: UCB1

set of arms A, number of rounds T;
for $t = 1, \dots, T$ **do**
    for $a \in A$ **do**
        $\mu_t(a) \leftarrow \frac{1}{N_{t-1}(a)} \sum_{t'=1}^{t-1} r_{t'}(a) \mathbb{I}[a_{t'} = a]$;
        $UCB_t(a) \leftarrow \mu_t(a) + \sqrt{\frac{2\log(T)}{N_{t-1}(a)}}$;
    play arm $a_t \in \arg\max_a UCB_t(a)$;

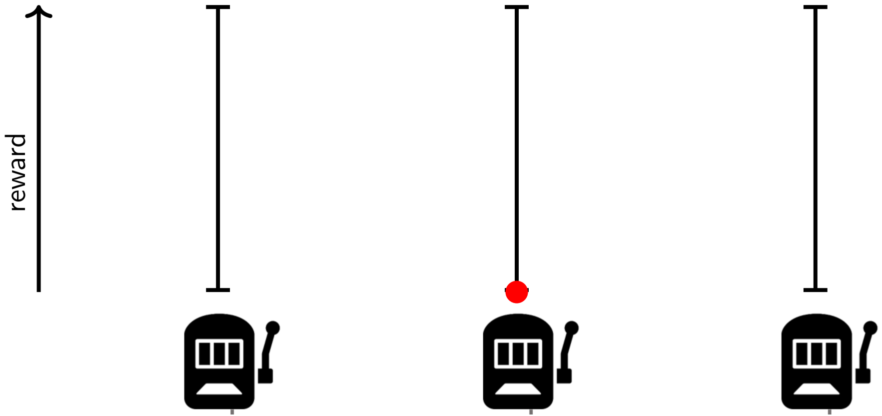
UCB1 example



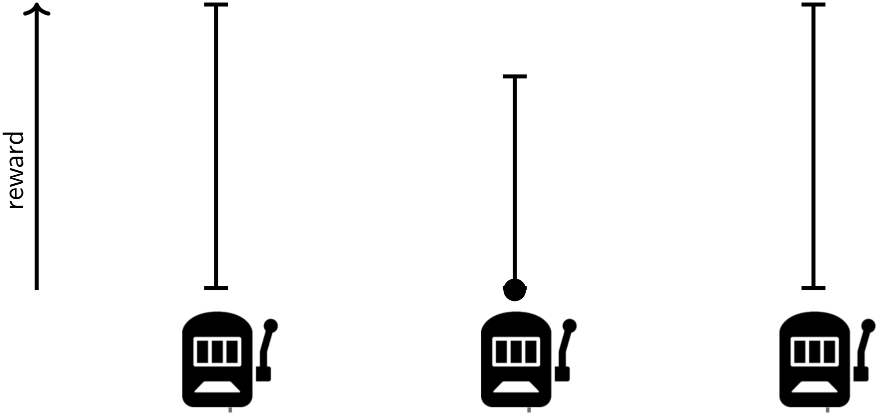
UCB1 example



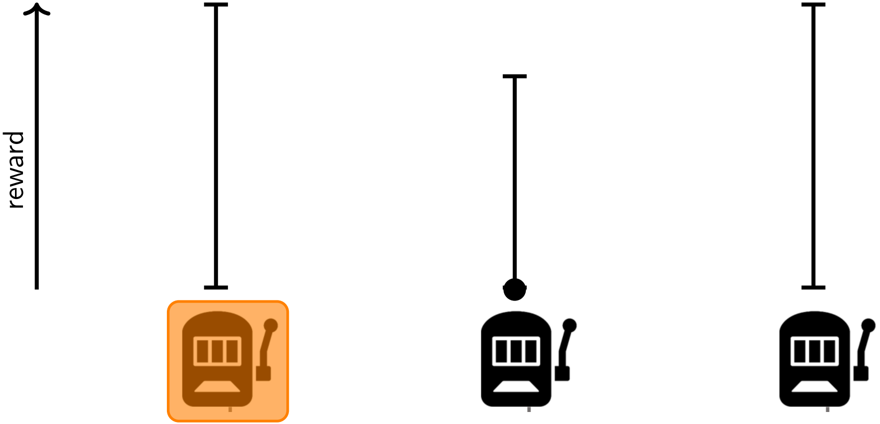
UCB1 example



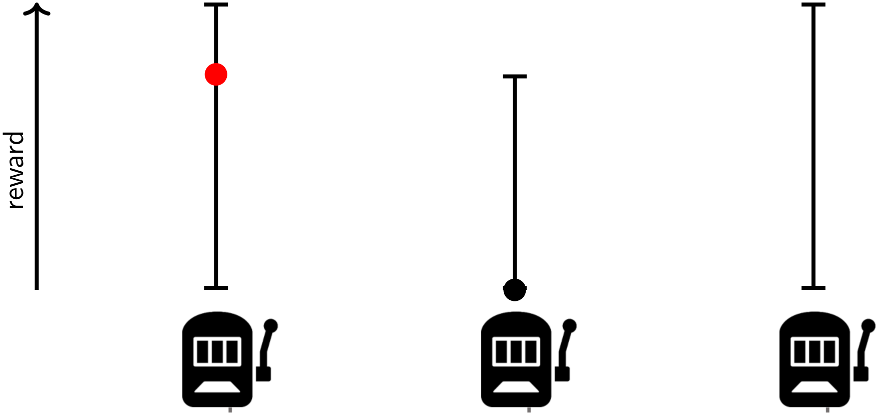
UCB1 example



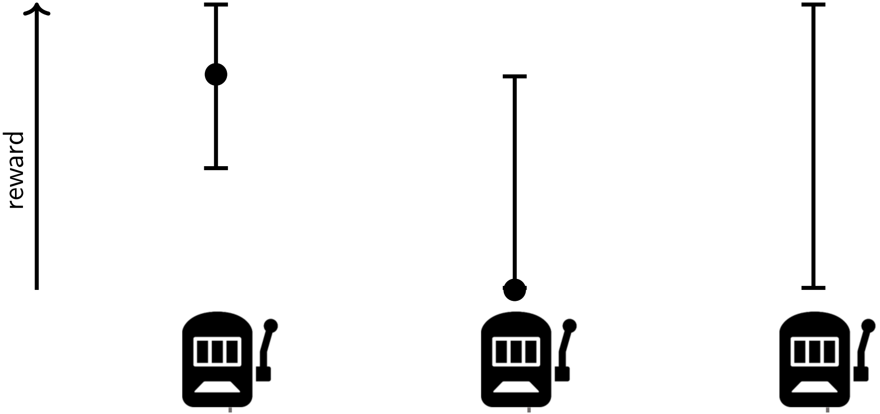
UCB1 example



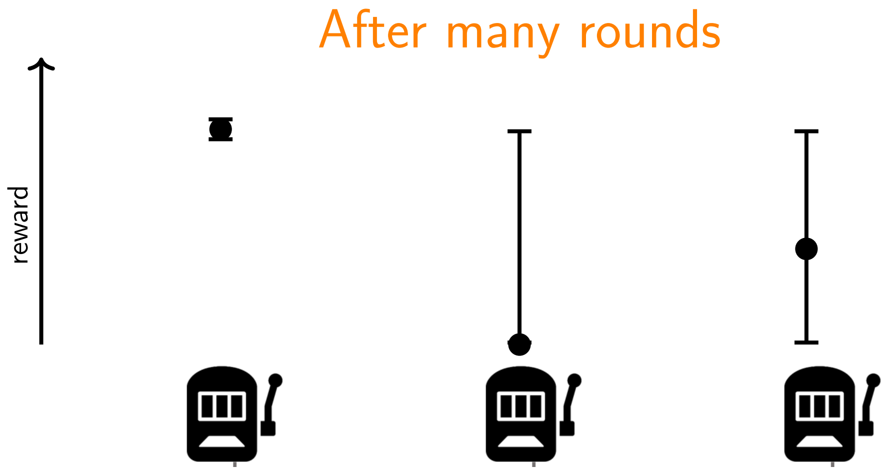
UCB1 example



UCB1 example

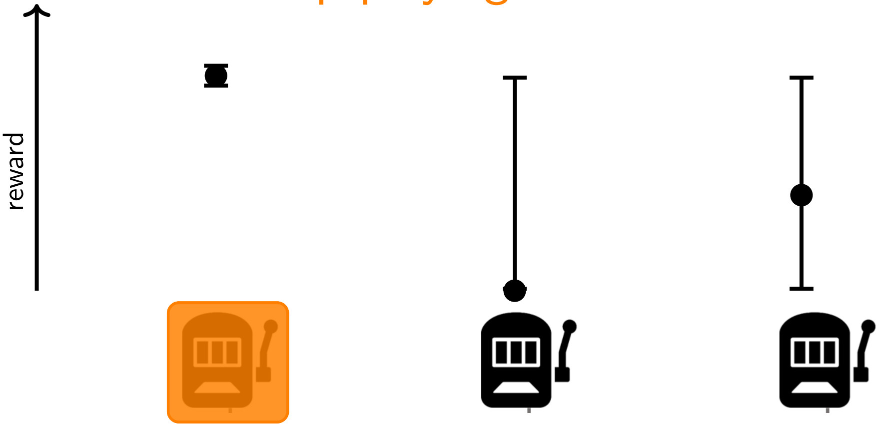


UCB1 example



UCB1 example

Keep playing the first arm



Clean event

Idea: The mean idea of UCB1 is that the empirical mean are “close” to the actual mean when an arm is played a sufficient number of times.

- This is true only with high probability
- We call **Clean Event** the high probability event in which all the estimations are close to the actual means

Lemma (Clean event)

$$\mathbb{P}\left(|\mu(a) - \mu_t(a)| \leq \sqrt{\frac{2\log(T)}{N_{t-1}(a)}} \quad \forall a, t\right) \geq 1 - 1/T.$$

Clean event

Idea: The mean idea of UCB1 is that the empirical mean are "close" to the actual mean when an arm is played a sufficient number of times.

- This is true only with high probability
- We call **Clean Event** the high probability event in which all the estimations are close to the actual means

Lemma (Clean event)

$$\mathbb{P}\left(|\mu(a) - \mu_t(a)| \leq \sqrt{\frac{2\log(T)}{N_{t-1}(a)}} \quad \forall a, t\right) \geq 1 - 1/T.$$

⚠ To prove the lemma we cannot use Hoeffding bound directly since the number of observations from each arm is random

- UCB1 is always **optimistic** about the means with high probability

How many times do we play a sub-optimal arm?

- We assume that the clean event holds
- We can bound the number of times a sub-optimal arm is played

Lemma

If the clean event holds, for any arm $a \neq a^*$:

$$N_T(a) \leq \frac{9 \log(T)}{\Delta_a^2}.$$

How many times do we play a sub-optimal arm?

Proof.

Let t be the last time we play arm a. Then,

$$\mu(a) + \sqrt{\frac{8\log(T)}{N_{t-1}(a)}} \geq \mu_t(a) + \sqrt{\frac{2\log(T)}{N_{t-1}(a)}} = UCB_t(a) \geq UCB_t(a^*) \geq \mu(a^*),$$

implying

$$N_{t-1}(a) \leq \frac{8\log(T)}{\Delta_a^2}.$$

The total number of pulls of arm a is at most

$$N_T(a) \leq \frac{8\log(T)}{\Delta_a^2} + 1 \leq \frac{9\log(T)}{\Delta_a^2}.$$

□

Regret Bound

Theorem

UCB1 achieves pseudo-regret $O(\log(T)\sum_{a\in A}\frac{1}{\Delta_a})$.

Proof.

- Assume that the clean event holds
  - ▷ Each arm $a \neq a^*$ is played at most $N_T(a) \leq \frac{9\log(T)}{\Delta_a^2}$ times
  - ▷ The total regret is at most $\sum_{a\in A}\Delta_a N_T(a) \leq 9\log(T)\sum_{a\in A}\frac{1}{\Delta_a}$
- Assume that the clean event does not hold
  - ▷ The regret is at most T

The expected regret is at most

$$(1-\frac{1}{T})9\log(T)\sum_{a\in A}\frac{1}{\Delta_a} + T\frac{1}{T} \leq O\left(\log(T)\sum_{a\in A}\frac{1}{\Delta_a}\right)$$

□

An instance-independent regret bound

- When $\Delta \to 0$ the regret goes to ∞
- We can derive a regret bound independent from Δ_a

Theorem

UCB1 achieves pseudo-regret:

$$\mathcal{R}_T = O(\sqrt{KT \log(T)}).$$

Proof Sketch.

Idea: If we play an arm that is $\sqrt{\frac{K\log(T)}{T}}$ sub-optimal it is fine.

- Consider the set $\hat{A}$ of arms a with $\Delta_a \leq \sqrt{\frac{K\log(T)}{T}}$:
  - The regret from playing arm $a \in \hat{A}$ is $\Delta_a N_T(a)$
  - The regret from such arms is at most $\sqrt{\frac{K\log(T)}{T}} \sum_{a\in\hat{A}} N_T(a) = \sqrt{KT\log(T)}$
- Consider the arms $a \notin \hat{A}$, i.e., with $\Delta_a > \sqrt{\frac{K\log(T)}{T}}$:
  - The regret from playing arm $a \notin \hat{A}$ is $\Delta_a N_T(a) \leq \Delta_a O\left(\frac{\log(T)}{\Delta_a^2}\right) \leq O\left(\sqrt{\frac{T\log(T)}{K}}\right)$
  - Since there are at most K of such arms, the regret from such arms is at most:

$$K \cdot O\left(\sqrt{\frac{T\log(T)}{K}}\right) \leq O\left(\sqrt{KT\log(T)}\right)$$

Thompson sampling

Thompson sampling

Another approach to tackle Stochastic MAB is the **Bayesian approach**:

- For every arm, have a prior distribution on its expected value
- For every arm, draw a sample according to the corresponding prior distribution
- Choose the arm with the best sample
- Update the prior distribution of the chosen arm according the observed realization

Beta distribution

We focus on Bernulli reward distributions (i.e., supported in $\{0,1\}$) and we use a Beta distribution as prior distribution.

Beta Distribution

The density of $Beta(\alpha, \beta)$ is defined as:

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1},$$

where $\Gamma(n) = (n-1)!$

Update the distribution

- We start with prior $Beta(1, 1) \rightarrow$ uniform distribution over $[0, 1]$
- When we observe a new sample we increase α if we observe $r_t(a) = 1$ or β if we observe $r_t(a) = 0$

$$(\alpha_a, \beta_a) \leftarrow (\alpha, \beta) + (r_t(a), 1 - r_t(a))$$

- The Beta distribution is a probability distribution over the expected value of the arm

Choose which arm to play

- For each arm a, sample $\theta_a \sim Beta(\alpha_a, \beta_a)$
- Play the arm with the largest sampled mean $a_t \in \arg\max \theta_a$

Thompson sampling

Algorithm: Thompson sampling

```
set of arms A, number of rounds T;
for a ∈ A do
    α_a = β_a = 1 ;
for t = 1, ..., T do
    for a ∈ A do
        θ_a ~ Beta(α_a, β_a) ;
    play arm a_t ∈ arg max θ_a;
    update (α_{a_t}, β_{a_t}) ← (α_{a_t}, β_{a_t}) + (r_t(a_t), 1 − r_t(a_t));
```

Thompson sampling

Similarly to UCB1, Thompson sampling provides $\log(T)$ instance-dependent regret.

Theorem

For each $\epsilon > 0$, Thompson sampling achieves pseudo-regret:

$$\mathcal{R}_T \leq (1+\epsilon) C \log(T) + C'/\epsilon^2,$$

where C and C' depend only on the reward functions.

- Similarly to UCB1, the constants C and C' can be arbitrarily large when Δ_a are small
- Usually Thompson sampling provides better empirical performances than UCB1

References

Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.