

Reinforcement Learning

Master IASD, Université PSL

<https://www.di.ens.fr/olivier.cappe/Courses/IASD-FoRL/>

November 2023



Table of Contents

- 1 The Multi-Armed Bandit Model and the Bayesian View
 - The Multi-Armed Bandit Model
 - Bayesian Algorithms
- 2 Fundamental Limits of Performance
- 3 Analysis of the UCB Algorithm
- 4 Other Bandit Algorithms

Bandits?



In the context of this course might be more accurately described as a **single-state MDP!**

A Short History of Bandits

- Thompson (1933) *On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples*. Biometrika
- Robbins (1952) *Some Aspects of the Sequential Design of Experiments*. Bull. Amer. Math. Soc.
- Gittins (1979) *Bandit Processes and Dynamic Allocation Indices*. J. R. Stat. Soc. Ser. B Stat. Methodol.
- Lai & Robbins (1985) *Asymptotically efficient adaptive allocation rules*. Adv. in Appl. Math. + Graves & Lai (1997) *Asymptotically Efficient Adaptive Choice of Control Laws in Controlled Markov Chains*. SIAM J. Control Optim.
- Auer, Cesa-Bianchi & Fischer (2002) *Finite-time Analysis of the Multi-Armed Bandit Problem*. Machine Learning Journal

And many papers since then in the machine learning literature...

The bandit model is an MDP with a single state; in this context, actions are usually referred to as "arms" (i.e. $\{A_t = a\}$ corresponds to "playing arm a ")

Definition (Multi-Armed Bandit Model)

A joint process $(A_t, X_t)_{t \geq 1}$, with $A_t \in \{1, \dots, k\}$ such that given the history $H_{t-1} = (A_1, X_1), \dots, (A_{t-1}, X_{t-1})$

- The agent may chose A_t as a function of H_{t-1} and, possibly, of an external independent randomization
- Given H_{t-1} and A_t , the environment generates X_t such that

$$\mu_a = \mathbb{E}[X_t | H_{t-1}, A_t = a] = \mathbb{E}[X_t | A_t = a] \quad (\text{arm-dependent expected reward})$$

When the parameters $\{\mu_a\}_{a=1, \dots, k}$ are known, the optimal action always consists in choosing $a_\star \in \operatorname{argmax}_{a \in \{1, \dots, k\}} \mu_a$

In the bandit model, eventually finding the best arm, and hence the best policy, is easy

For instance, asynchronous Q-learning with $\alpha_t = 1/t$

- ① Explores all arms with some stochastic policy π
- ② Estimates the unknown parameters μ_1, \dots, μ_k according to

$$\hat{\mu}_a(t) = \frac{S_a(t)}{N_a(t)}$$

where

- $S_a(t) = \sum_{i=1}^t X_i \mathbb{1}\{A_i = a\}$
- $N_a(t) = \sum_{i=1}^t \mathbb{1}\{A_i = a\}$

(note that $N_a(t)$ is random)

We will be interested in algorithms that do this optimally

Bayesian Optimal Algorithms

In this part we assume a **prior** distribution λ on the set of bandit problems (on the reward distributions v_1, \dots, v_k associated to the k arms), and consider the **Bayesian reward**

$$\mathbb{E}_{(v_1, \dots, v_k) \sim \lambda} \left(\mathbb{E} \left[\sum_{t \geq 0} X_t \middle| v_1, \dots, v_k \right] \right)$$

- 1 The inner conditional expectation averages w.r.t. the random rewards X_t and, possibly, the internal randomization of the bandit algorithm
- 2 The outer expectation averages over all possible models (under λ)

Interestingly, the Bayesian framework makes it possible to define optimal bandit algorithms using MDP ideas (which are however not practical)

The Bayesian Approach

By specifying a prior distribution λ on an unknown parameter θ , the knowledge on θ gained from observing X_1, \dots, X_t is fully summarized by the **posterior distribution**

$$\Lambda_t(\theta) = p(\theta|X_1, \dots, X_t) = \frac{p(X_1, \dots, X_t|\theta)\lambda(\theta)}{\int p(X_1, \dots, X_t|\theta')\lambda(\theta')d\theta'}$$

which defines

Posterior mean estimator $\int \theta \Lambda_t(\theta) d\theta$

Predictive distribution $\int p(x_{t+1}|X_1, \dots, X_t, \theta) \Lambda_t(\theta) d\theta$

Posterior probability of hypothesis $\theta \in \mathcal{R}$ $\int_{\mathcal{R}} \Lambda_t(\theta) d\theta$

Sequential update $\Lambda_{t+1}(\theta) \propto p(X_{t+1}|X_1, \dots, X_t, \theta) \Lambda_t(\theta)$



Bayesian computation are usually not available in closed-form, except when using **conjugate priors**

Example: Beta – Binomial Bayesian Experiment

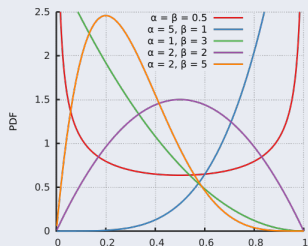
Prior $\theta \sim \text{Beta}(\alpha, \beta)$, Likelihood $X_i|\theta \sim \text{Bernoulli}(\theta)$ (with X_1, X_2, \dots i.i.d given θ)

Definition (Beta Distribution)

$$\text{PDF } \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

$$\text{Expectation } m = \frac{\alpha}{\alpha + \beta}$$

$$\text{Variance } \frac{m(1-m)}{\alpha + \beta + 1} \leq \frac{1}{4(\alpha + \beta + 1)}$$



- Posterior $\theta|X_1, \dots, X_n \sim \text{Beta}(\alpha + S_n, \beta + n - S_n)$, where $S_n = \sum_{i=1}^n X_i$
- Predictive distribution $X_{n+1}|X_1, \dots, X_n \sim \text{Bernoulli}\left(\frac{\alpha + S_n}{\alpha + \beta + n}\right)$



Bayesian Optimal Bandit as an MDP Planning Problem

Denote by $\Lambda_t = (\Lambda_{1,t}, \dots, \Lambda_{k,t})$ the posterior distributions at time t of the unknown parameters of the arm-dependent distributions ν_1, \dots, ν_k

Letting $S_t = \Lambda_{t-1}$ makes it possible to view the bandit Bayesian learning problem as an MDP where, given S_t and A_t ,

- $\mathbb{E}[X_t | S_t, A_t = a] = \int x \int \nu_a(x|\theta) \Lambda_{a,t-1}(\theta) d\theta dx$
(expectation of the predictive distribution)
- $\Lambda_{a,t}(\theta | S_t, A_t = a) \propto \nu_a(X_t|\theta) \Lambda_{a,t-1}(\theta)$
where X_t is distributed under the predictive distribution $\int \nu_a(x|\theta) \Lambda_{a,t-1}(\theta) d\theta$
(sequential posterior update)

When using conjugate priors, Λ_t can be represented by finite-dimensional statistics, e.g. for Bernoulli bandits $\Lambda_{a,t} = \text{Beta}(\alpha_a + S_a(t), \beta_a + (N_a(t) - S_a(t)))$

Bayesian Optimal Bandit as an MDP Planning Problem

The Bayesian optimal learning algorithm may be found as the solution of the planning problem

Let's see how it works in

- Two armed bandit ($k=2$)
- With Bernoulli distributions ($\mathbb{P}(X_a = 1) = \mu_a$, $\mathbb{P}(X_a = 0) = 1 - \mu_a$)
- For horizon $n=2$ (homework: do it for $n=3$ at home...)
- Using independent $\text{Beta}(\alpha_a, \beta_a)$ priors on μ_a



\mathcal{H}_1		arm posteriors		$q_{1:2}(\mathcal{H}_1, A_2=1)$	$q_{1:2}(\mathcal{H}_1, A_2=2)$
A_1	x_1				
1	1	$\alpha_{1,1} = \alpha_{1,0} + 1$ $\beta_{1,1} = \beta_{1,0}$	$\alpha_{2,1} = \alpha_{2,0}$ $\beta_{2,1} = \beta_{2,0}$	$\frac{\alpha_{1,0} + 1}{\alpha_{1,0} + \beta_{1,0} + 1}$	$\frac{\alpha_{2,0}}{\alpha_{2,0} + \beta_{2,0}}$
1	0	$\alpha_{1,1} = \alpha_{1,0}$ $\beta_{1,1} = \beta_{1,0} + 1$	$\alpha_{2,1} = \alpha_{2,0}$ $\beta_{2,1} = \beta_{2,0}$	$\frac{\alpha_{1,0}}{\alpha_{1,0} + \beta_{1,0} + 1}$	$\frac{\alpha_{2,0}}{\alpha_{2,0} + \beta_{2,0}}$
2	1	same thing with arm 1 \leftrightarrow arm 2			
2	0				

$$q_{1:2}(A_1 = 1) = \frac{\alpha_{1,0}}{\alpha_{1,0} + \beta_{1,0}} + \frac{\alpha_{1,0}}{\alpha_{1,0} + \beta_{1,0}} \max \left(\frac{\alpha_{1,0} + 1}{\alpha_{1,0} + \beta_{1,0} + 1}, \frac{\alpha_{2,0}}{\alpha_{2,0} + \beta_{2,0}} \right) + \frac{\beta_{1,0}}{\alpha_{1,0} + \beta_{1,0}} \max \left(\frac{\alpha_{1,0}}{\alpha_{1,0} + \beta_{1,0} + 1}, \frac{\alpha_{2,0}}{\alpha_{2,0} + \beta_{2,0}} \right)$$

Gittins Indices

In the infinite horizon, γ -discounted case, Gittins (1979) showed that the optimal policy is an **index policy** where, if $\Lambda_{a,t-1}$ denotes the posterior on arm a at time $t-1$

$$A_t = \operatorname{argmax}_{a \in \{1, \dots, k\}} g_\gamma(\Lambda_{a,t-1})$$

Definition (Gittins index)

$$g_\gamma(\lambda) = \inf \left\{ \rho : \sup_{T \geq 0} \mathbb{E}_\lambda \left[\sum_{t=1}^T \gamma^{t-1} X_t + \frac{\gamma^T \rho}{1-\gamma} \right] = \frac{\rho}{1-\gamma} \right\}$$

where the supremum is taken over all **random stopping times** T^*

$g_\gamma(\lambda)$ can be interpreted as the exploration threshold in the one-armed bandit model with retirement (with prior λ on the unknown arm)



*

A stopping time is a random variable such that $\mathbb{1}_{\{T=t\}}$ may be expressed as a function of X_1, \dots, X_t

Thompson Sampling

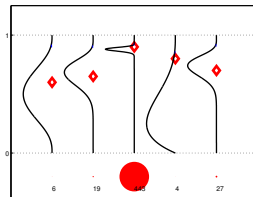
A Bayesian-inspired randomized algorithm that is successfully used in practice, has been proposed by Thompson (in 1933!) but was only analyzed very recently

Thompson Sampling

- Draw $I_{a,t}$ from each posterior distribution $\Lambda_{a,t-1}$, for $a = 1, \dots, k$
- Select

$$A_t = \arg \max_{a \in \{1, \dots, k\}} I_{a,t}$$

- Observe X_t and update the posterior $\Lambda_{A_t, t-1}$ to obtain $\Lambda_{A_t, t}$



A key observation is that Thompson sampling selects arm a according to the Bayesian posterior probability $\mathbb{P}(A_\star = a | H_{t-1})$ that it is actually optimal

Thompson Sampling for Bernoulli with Beta Priors

Thompson Sampling for Binary Rewards

- Initialize $(\alpha_{a,0}, \beta_{a,0}) = (\alpha, \beta)$ for $a = 1, \dots, k$
- For $t = 1, \dots$
 - Draw $I_{a,t} \sim \text{Beta}(\alpha_{a,t-1}, \beta_{a,t-1})$ for $a = 1, \dots, k$
 - Select

$$A_t = \underset{a \in \{1, \dots, k\}}{\operatorname{argmax}} I_{a,t}$$

- Observe X_t and update the posterior parameter for arm A_t :

$$\begin{cases} \alpha_{A_t,t} &= \alpha_{A_t,t-1} + X_t \\ \beta_{A_t,t} &= \beta_{A_t,t-1} + (1 - X_t) \end{cases}$$

A good exercise is to derive the Thompson sampling algorithm for Gaussian rewards (with known variance) assuming centered Gaussian prior on the μ_a parameters

Table of Contents

- ① The Multi-Armed Bandit Model and the Bayesian View
- ② Fundamental Limits of Performance
 - The Regret Decomposition
 - Minimax Lower Bound
 - The Lai and Robbins Lower Bound
- ③ Analysis of the UCB Algorithm
- ④ Other Bandit Algorithms

Recall that in the **bandit model**

$$\mathbb{E}[X_t | A_t = a, H_{t-1}] = \mu_a$$

and that the best action is

$$a_\star \in \operatorname{argmax}_{a \in \{1, \dots, k\}} \mu_a$$

We will also use the notations

- $\mu_\star = \max_{a \in \{1, \dots, k\}} \mu_a$
- $\Delta_a = \mu_\star - \mu_a$ for the expected reward gap between a and an optimal action a_\star
- ν_a for the conditional distribution of X_t given $A_t = a$ and H_{t-1}

w.l.o.g. we will also assume in our notations that a_\star is unique

From Reward Maximization to Regret Minimization

The usual criterion considered in this context is the **expected regret**

Definition ((Stochastic) Regret or Pseudo-Regret)

$$R_n = \max_{a \in \{1, \dots, k\}} \mathbb{E} \left[\sum_{t=1}^n X_t \middle| A_1 = \dots = A_n = a \right] - \sum_{t=1}^n X_t$$

Proposition

$$\mathbb{E}[R_n] = n\mu_{\star} - \mathbb{E} \left[\sum_{t=1}^n X_t \right] = \sum_{\substack{a=1 \\ a \neq a_{\star}}}^k \Delta_a \mathbb{E}[N_a(n)]$$

where $N_a(n) = \sum_{t=1}^n \mathbb{1}\{A_t = a\}$



↪ A **sequential decision** task that is not equivalent to estimating the values of the arm means (μ_a)

Naive Strategies Have Linear Regret

- Draw a Fixed Arm a

$$\mathbb{E}(R_n) = n\Delta_a$$

- Draw Random Arm

$$\mathbb{E}(R_n) = \frac{n}{k} \sum_{\substack{a=1 \\ a \neq a_\star}}^k \Delta_a$$

But also

- Epsilon-Greedy with Fixed ϵ

$$\mathbb{E}(R_n) \geq \epsilon \frac{n}{k} \sum_{\substack{a=1 \\ a \neq a_\star}}^k \Delta_a$$

- Explore Then Commit with Fixed or Proportional Exploration Time m

$$\mathbb{E}(R_n) = \frac{m}{k} \sum_{\substack{a=1 \\ a \neq a_\star}}^k \Delta_a + (n-m) \sum_{\substack{a=1 \\ a \neq a_\star}}^k \Delta_a \mathbb{P}(A_{m+1} = a)$$

Alternative Objective: Best Arm Identification

Goal: Find which of the k hypotheses $\mathcal{H}_a: \mu_a > \mu_b$, for $b \neq a$, is true

Definition (Fixed Confidence Setting)

Given a probability δ , design an allocation rule and a stopping time T such that

- ① $\mathbb{P}(A_{T+1} \neq k^*) < \delta$
- ② $\mathbb{E}[T]$ is minimal

- related to classical sequential hypothesis testing, with active added allocation
- requires more exploration than the reward maximization objective
- will not be addressed in the rest of this course

Theorem (Data-Processing Inequality)

Let P and Q be two probability measures and Z a random variable

$$\text{KL}(P, Q) \geq \text{KL}(P^Z, Q^Z)$$

where P^Z (resp. Q^Z) denote the distribution of Z under P (resp. under Q)

Corollary (Garivier, Ménard & Stoltz, 2018, arXiv:1602.07182)

For any random variable $Z \in [0, 1]$,

$$\text{KL}(P, Q) \geq d(\mathbb{E}_P[Z], \mathbb{E}_Q[Z])$$

where d is the Bernoulli Kullback-Leibler divergence

$$d(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1-p}{1-q}$$



Lemma (KL Divergence between Two MAB Models)

Consider two stochastic MAB models with arm distributions ν and ν' ; for any algorithm one has

$$\text{KL}(\mathbb{P}_{\nu}^{X_1, \dots, X_n}, \mathbb{P}_{\nu'}^{X_1, \dots, X_n}) = \sum_{a=1}^k \text{KL}(\nu_a, \nu'_a) \mathbb{E}_{\nu}[N_a(n)]$$

In particular, if $\nu = (\nu_1, \dots, \nu_a, \dots, \nu_k)$ and $\nu' = (\nu_1, \dots, \nu'_a, \dots, \nu_k)$,

$$\text{KL}(\mathbb{P}_{\nu}^{X_1, \dots, X_n}, \mathbb{P}_{\nu'}^{X_1, \dots, X_n}) = \text{KL}(\nu_a, \nu'_a) \mathbb{E}_{\nu}[N_a(n)]$$



Proof Scheme

The two inequalities that follow are obtained using the same idea:

- ① For a bandit model ν , chose an arm a that is sub-optimal, that is, such that $\mu_a < \mu_\star$
- ② Consider the model ν' in which all arms remain unchanged, except arm a for which $\mu'_a > \mu_\star$ (i.e. under ν' , a becomes the optimal arm)
- ③ Use the inequality

$$\text{KL}(\nu_a, \nu'_a) \mathbb{E}_\nu[N_a(n)] \geq d\left(\frac{\mathbb{E}_\nu[N_a(n)]}{n}, \frac{\mathbb{E}_{\nu'}[N_a(n)]}{n}\right)$$

Two Useful Facts About Kullback-Leibler Divergences

Lemma ((Basic) Pinsker Inequality)

$$d(p, q) \geq 2(p - q)^2$$



Lemma (Gaussian KL)

$$\text{KL}(\mathcal{N}(\mu, \sigma^2), \mathcal{N}(\mu', \sigma^2)) = \frac{(\mu - \mu')^2}{2\sigma^2}$$



Minimax Lower Bound

Theorem (Minimax Gaussian Two Arm Lower Bound)

Consider a bandit model with two unknown Gaussian $\mathcal{N}(\mu_1, \sigma^2), \mathcal{N}(\mu_2, \sigma^2)$ arms

For any algorithm and horizon n , there exist instances of the model for which

$$\mathbb{E}[R_n] \geq \frac{\sigma}{8} \sqrt{n}$$

For k -armed bandits, similar results hold with a lower bound of the form $c\sqrt{kn}$
[Lattimore & Szepesvári, 2020; Chap. 15]

- We are now interested in lower bound results that pertain to a specific instance of the model (unfortunately these will be asymptotic, i.e., requiring that $n \rightarrow \infty$)
- These bounds can hold only for algorithms that are efficient for all model instances

Definition (Consistent Strategy)

A strategy is consistent if for any parameters ν of the stochastic MAB model and all $\alpha > 0$,

$$\lim_{n \rightarrow \infty} \frac{\mathbb{E}_\nu[R_n]}{n^\alpha} = 0$$

This implies that for all $a \neq a_\star$,

$$\lim_{n \rightarrow \infty} \frac{\mathbb{E}_\nu[N_a(n)]}{n^\alpha} = 0$$

Proposition

Consider a MAB model in which v is parameterized by its expectation (i.e., for any feasible value of $\mu \in \mathbb{R}$, there is a unique v which has expectation μ)

For any consistent strategy and $a \neq a_\star$, and under regularity conditions,

$$\liminf_{n \rightarrow \infty} \frac{\mathbb{E}_v[N_a(n)]}{\log n} \geq \frac{1}{\text{KL}(v_a, v_\star)}$$

Corollary (Lai and Robbins Lower Bound)

In the same conditions,

$$\liminf_{n \rightarrow \infty} \frac{\mathbb{E}_v[R_n]}{\log n} \geq \sum_{\substack{a=1 \\ a \neq a_\star}}^k \frac{\Delta_a}{\text{KL}(v_a, v_\star)}$$


For Gaussian MAB, the r.h.s. equals $\sum_{a \neq a_\star} \sigma^2 / (2\Delta_a)$

Proof Hint

- Given v , consider any sub-optimal arm $a \neq a_\star$, i.e., such that $\mu_a < \mu_\star$
- Choose v' such that v'_a is the best arm, i.e., $\mu'_a > \mu_\star$, while all other arms but a remain unchanged

As before, apply our main inequality to $Z = N_a(n)/n$ and use the inequality

$$d(p, q) \geq (1 - p) \log \frac{1}{1 - q} - \log 2$$

Let n tend to infinity and μ'_a tend to μ_\star (from above) 

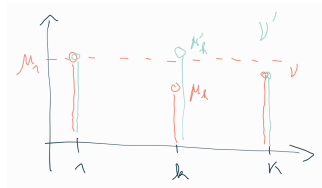


Table of Contents

- ① The Multi-Armed Bandit Model and the Bayesian View
- ② Fundamental Limits of Performance
- ③ Analysis of the UCB Algorithm
 - Deviation Inequalities
 - The Explore-Then-Commit (ETC) Algorithm
 - The Upper Confidence Bound (UCB) Algorithm
- ④ Other Bandit Algorithms

Why Do We Need Deviation Inequalities?

Contrary to deterministic or purely randomized allocations, bandit allocation does not preserve distributions: the distribution of $S_a(t)$ (or even $S_a(t)|N_a(t) = n$) depends on the algorithm

The following is true:

- $X_t|H_{t-1}, A_t = a \sim \nu_a$
- In the Bayesian approach, if a conjugate prior distribution λ is specified on (ν_k) , the posterior distribution, given H_{t-1} is also available in closed form
-

$$S_a(t) - \mu_a N_a(t) \quad \text{is a martingale}$$

implying, in particular, that

$$\mathbb{E}[S_a(t)] = \mu_a \mathbb{E}[N_a(t)]$$



(we will not use this more advanced argument in this course)

Typical Use of Deviation Inequalities (in this course)

Imagine that

- ① behind each arm there are infinite sources $(X_{a,i})_{i \geq 1}$ of i.i.d. random variables of distribution ν_a
- ② and that when $A_t = a$, $X_t = X_{A_t, N_{A_t}(t)}$

We need a **maximal deviation inequality** (deviation for the supremum):

$$\begin{aligned} \mathbb{P}\left(\sqrt{N_a(t)}(\bar{X}_a(t) - \mu_a) > \delta\right) &\leq \mathbb{P}\left(\max_{1 \leq m \leq t} \sqrt{m}(\bar{X}_{a,m} - \mu_a) > \delta\right) \\ &= \mathbb{P}\left(\exists m, 1 \leq m \leq t: \sqrt{m}(\bar{X}_{a,m} - \mu_a) > \delta\right) \leq \sum_{m=1}^t \mathbb{P}\left(\sqrt{m}(\bar{X}_{a,m} - \mu_a) > \delta\right) \end{aligned}$$

(union bound)

where $\bar{X}_a(t) = 1/N_a(t) \sum_{s=1}^t X_s \mathbb{1}\{A_s = a\}$ and $\bar{X}_{a,m} = 1/m \sum_{i=1}^m X_{a,i}$

Lemma (Cramér-Chernoff Method)

Assume $(X_i)_{i \geq 1}$ i.i.d. $\sim \nu$, with $\mathbb{E}[e^{\lambda X_1}] < \infty$, $\forall \lambda \in \mathbb{R}$. Let $\mu = \mathbb{E}[X_1]$, $\bar{X}_n = 1/n \sum_{i=1}^n X_i$, $\phi(\lambda) = \log \mathbb{E}[e^{\lambda X_1}]$ and $I(x) = \phi^*(x) = \sup_{\lambda \in \mathbb{R}} \lambda x - \phi(\lambda)$. For $x > \mu$,

$$\mathbb{P}(\bar{X}_n > x) \leq e^{-nI(x)}$$



Lemma (Underestimation Bound)

Under the same conditions, for $x < \mu$,

$$\mathbb{P}(\bar{X}_n < x) \leq e^{-nI(x)}$$

These results are non improvable “in rate”, in the sense of the following Large Deviation Theorem

Theorem (Cramér Theorem)

Under the same conditions,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(\bar{X}_n > x) = -I(x) \quad (x > \mu)$$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(\bar{X}_n < x) = -I(x) \quad (x < \mu)$$

Lemma (Gaussian Deviation Bound (Underestimation))

If $X_1 \sim \mathcal{N}(\mu, \sigma^2)$, $\phi(\lambda) = \lambda^2 \sigma^2 / 2 + \mu \lambda$, $I(x) = (x - \mu)^2 / (2\sigma^2)$.

Hence, for $x < \mu$,

$$\mathbb{P}(\bar{X}_n < x) \leq e^{-n \frac{(x - \mu)^2}{2\sigma^2}}$$



Corollary (Gaussian Upper Confidence Bound)

For any probability $\delta \in (0, 1)$,

$$\mathbb{P}\left(\bar{X}_n + \sqrt{\frac{2\sigma^2}{n} \log \frac{1}{\delta}} < \mu\right) \leq \delta$$



Lemma (Hoeffding Lemma)

If $X_1 \in [0, 1]$, $\phi(\lambda) \leq \lambda^2/8 + \mu\lambda$, i.e., “ v is $1/2$ — sub-Gaussian”



Thus for $X_1 \in [0, 1]$, the previous bounds hold with $\sigma^2 = 1/4$, in particular,

$$\mathbb{P}\left(\bar{X}_n + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}} < \mu\right) \leq \delta$$

Warning: Assuming that rewards are in $[0, 1]$ is the most common assumption in the bandit literature (used in this course) but others —such as, e.g., Lattimore & Szepesvári’s book— consider instead 1-sub-Gaussian rewards

Useful additional result

Lemma (Lemma 5.4 of Lattimore & Szepesvári)

If X_1 and X_2 are independent random variables that are, respectively, σ_1 and σ_2 sub-Gaussian,

$$X_1 + X_2 \text{ is } \sqrt{\sigma_1^2 + \sigma_2^2} \text{ — sub-Gaussian}$$



The Basic Approach

Algorithm (Explore-then-Commit (ETC))

- For “rounds” $i = 1, \dots, m$, play arms $a = 1, \dots, k$ such that $N_a(mk) = m$ for each $a \in \{1, \dots, k\}$.
- For $t \geq 1 + mk$ play

$$A_t = \arg \max_{a \in \{1, \dots, k\}} \bar{X}_a(mk)$$

where

$$\bar{X}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^t X_s \mathbb{1}\{A_s = a\}$$

Note that by construction in ETC, the number of arm draws when choosing the arm for the second phase is deterministic

Theorem (Regret of ETC)

The regret of the Explore-Then-Commit algorithm may be bounded as

$$\mathbb{E}[R_n] \leq \sum_{\substack{a=1 \\ a \neq a_\star}}^k \Delta_a \left(m + ne^{-m\Delta_a^2} \right)$$



- The interesting regime occurs when $1 \ll m \ll n$
- Optimizing m requires knowledge of n and $\Delta_{\min} \leq (\Delta_a)$
— not anytime, not adaptive!
- The latter is very conservative

Instance (or Parameter) Dependent Bound

Taking $m = \left\lceil \frac{\log n}{\Delta_{\min}^2} \right\rceil$,

$$\mathbb{E}[R_n] \leq \sum_{\substack{a=1 \\ \neq a_\star}}^k \Delta_a \left(1 + \frac{\log n}{\Delta_{\min}^2} \right)$$



Minimax Bound

When $k=2$, taking $m = \left\lceil \frac{\log(n\Delta^2)}{\Delta^2} \right\rceil$ if $\Delta > \frac{1}{\sqrt{n}}$ and anything otherwise,

$$\mathbb{E}[R_n] \leq \sqrt{n}(1 + \log n)$$



In simple cases, ETC can be made adaptive (but not anytime)

Algorithm (Adaptive ETC (Two Arms))

Given an horizon n ,

- *Play arms 1 and 2, set $M = 2$*
- *While $|\bar{X}_1(2M) - \bar{X}_2(2M)| \leq \sqrt{\frac{\gamma \log n}{M}}$*
 - *Play arms 1 and 2*
 - *$M++$*
- *For $1 + 2M \leq t \leq n$*
 - *play $A_t = 1$ if $\bar{X}_1(2M) > \bar{X}_2(2M)$*
 - *or $A_t = 2$ otherwise*

Proposition

For $\gamma > 2$, Adaptive ETC satisfies

$$\mathbb{E}[R_n] \leq \frac{\gamma(1+\epsilon)\log n}{\Delta} + O_{\gamma,\epsilon}(1)$$

for all $\epsilon > 0$, where Δ denotes the gap between the two arms

Proof Hint

$$\mathbb{E}(R_n) \leq \Delta \mathbb{E}(M) + n \sum_{m=1}^{n/2} \mathbb{P} \left(\bar{X}_{1,m} - \bar{X}_{2,m} < -\sqrt{\frac{\gamma \log n}{m}} \right)$$

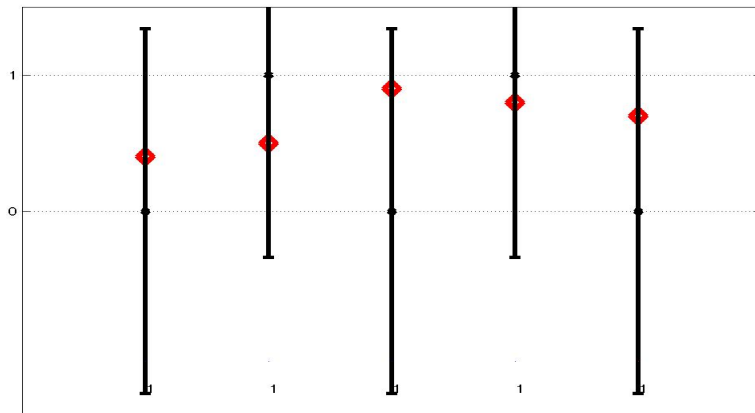


Algorithm (UCB Algorithm)

- For $t = 1, \dots, k$, play arms $1, \dots, k$
- For $t \geq k + 1$, Play $A_t = \operatorname{argmax}_{a \in \{1, \dots, k\}} U_a(t)$, where

$$U_a(t) = \bar{X}_a(t-1) + \sqrt{\frac{\gamma \log t}{2N_a(t-1)}}$$

UCB in Action



Proposition

For $\gamma > 2$, the regret of the UCB algorithm is bounded by

$$\mathbb{E}[R_n] \leq \sum_{\substack{a=1 \\ a \neq a_\star}}^k \frac{\gamma(1+\epsilon)\log n}{2\Delta_a} + O_{\gamma,\epsilon}(1)$$

for all $\epsilon > 0$

Note that using a more precise maximal inequality for $(\bar{X}_{k,t})_{t \geq 1}$ shows that the result also holds true for $\gamma \in (1, 2]$

Proof hint

Assume w.l.o.g. that $a_\star = 1$, and consider a suboptimal arm $a \neq 1$

$$N_a(n) \leq \sum_{t=1}^n \mathbb{1}\{U_1(t) < \mu_1\} + \sum_{t=1}^n \mathbb{1}\{U_a(t) > \mu_1, A_t = a\}$$



Table of Contents

- ① The Multi-Armed Bandit Model and the Bayesian View
- ② Fundamental Limits of Performance
- ③ Analysis of the UCB Algorithm
- ④ Other Bandit Algorithms
 - Thompson Sampling
 - The KL-UCB Algorithm
 - Contextual Bandits and the Lin-UCB Algorithm

Thompson Sampling

- Draw $I_{a,t}$ from each posterior distribution $\Lambda_{a,t-1}$, for $a = 1, \dots, k$
- Select

$$A_t = \operatorname{argmax}_{a \in \{1, \dots, k\}} I_{a,t}$$

- Observe X_t and update the posterior $\Lambda_{A_t,t-1}$ to obtain $\Lambda_{A_t,t}$

Satisfies

$$\mathbb{E}_{(v_1, \dots, v_k) \sim \lambda} (\mathbb{E}[R_n | v_1, \dots, v_k]) \leq O\left(\sqrt{kn \log n}\right)$$

A much more involved result is that it also satisfies the same instance-dependent bound on the non-Bayesian regret as KL-UCB

Proof hint

[Lattimore & Szepesvari, Th. 36.1] First note that

$$\mathbb{E}[\mu_{A_\star} - \mu_{A_t}] = \mathbb{E}[\mu_{A_\star} - U_{A_\star}(t)] + \mathbb{E}[U_{A_t}(t) - \mu_{A_t}]$$

And consider

$$U_a(t) = \bar{X}_a(t-1) + \sqrt{\frac{\log(1/\delta)}{2N_a(t-1)}}$$

Chernoff Bound for Bernoulli Distributions

Lemma (Bernoulli Concentration Bound (Underestimation))

If $X_1 \sim \text{Bernoulli}(\mu)$, $\phi(\lambda) = \log(1 + \mu(e^\lambda - 1))$, $I(x) = d(x, \mu)$, and, for $x < \mu$,

$$\mathbb{P}(\bar{X}_n < x) \leq e^{-nd(x, \mu)}$$



Corollary (Bernoulli Upper Confidence Bound)

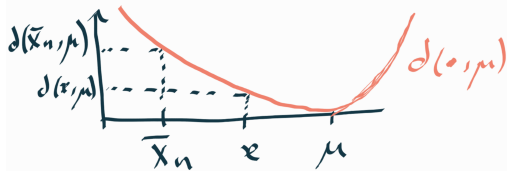
For any probability $\delta \in (0, 1)$,

$$\mathbb{P}\left(\bar{X}_n < \mu, n d(\bar{X}_n, \mu) > \log \frac{1}{\delta}\right) \leq \delta$$

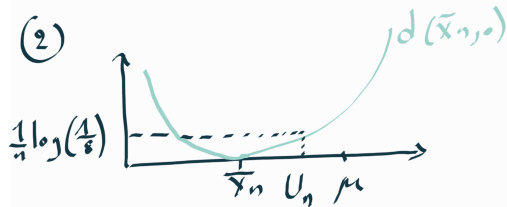
And $U_n \in (\bar{X}_n, 1)$ defined by $n d(\bar{X}_n, U_n) = \log(1/\delta)$ satisfies $\mathbb{P}(\mu > U_n) \leq \delta$

(1) If $\bar{x}_n < x < \mu$

$$\mathbb{P}(\bar{X}_n < x) = \mathbb{P}(d(\bar{X}_n, \mu) > d(x, \mu))$$



(2)



$$\mu > U_n \Leftrightarrow n d(\bar{X}_n, \mu) > \log \frac{1}{\delta}$$

Algorithm (KL-UCB Algorithm)

- For $t = 1, \dots, k$, play arms $1, \dots, k$
- For $t \geq k+1$, Play $A_t = \operatorname{argmax}_{a \in \{1, \dots, k\}} U_a(t)$, where $U_a(t) \in (\bar{X}_a(t-1), 1)$ is the solution of

$$N_a(t-1) d(\bar{X}_a(t-1), U_a(t)) = \gamma \log(t)$$

Proposition (Regret of KL-UCB)

Assuming Bernoulli rewards, the regret of the KL-UCB algorithm satisfies, for $\gamma > 1$,

$$\mathbb{E}[R_n] \leq \sum_{\substack{a=1 \\ a \neq a_\star}}^k \frac{\gamma(1+\epsilon)\Delta_a}{d(\mu_a, \mu_\star)} + O_{\gamma,\epsilon}(1)$$

for all $\epsilon > 0$

Linear Bandit Model

At time t , the learner has access to a (possibly) time-dependent set of actions $\mathcal{A}_t = \{a_{t,1}, \dots, a_{t,k}\} \in \mathbb{R}^d$ (with $\|a_{t,j}\|_2 \leq L$)

They can only be **probed one at a time**, i.e., the learner

- Chooses an action $A_t \in \mathcal{A}_t$
- and observes only the noisy **linear reward** $X_t = A_t^\top \theta^\star + \eta_t$ where η_t is a σ -subgaussian random noise (and $\|\theta^\star\|_2 \leq S$)

Note that when $\mathcal{A}_t = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix}$ the model reduces to the standard multiarmed bandit model

Optimality Criteria

Dynamic Regret Minimization

$$\begin{aligned}\max \mathbb{E} \left(\sum_{t=1}^n X_t \right) &\iff \min \mathbb{E} \left[\sum_{t=1}^n \max_{a \in \mathcal{A}_t} \langle a, \theta^* \rangle - \sum_{t=1}^n X_t \right] \\ &\iff \min \underbrace{\mathbb{E} \left(\sum_{t=1}^n \max_{a \in \mathcal{A}_t} \langle a - A_t, \theta^* \rangle \right)}_{\text{dynamic regret}}\end{aligned}$$

(Regularized) Least Squares Estimator

$$\hat{\theta}_t = \arg \min_{\theta \in \mathbb{R}^d} \sum_{s=1}^t (X_s - A_s^\top \theta)^2 + \lambda \|\theta\|_2^2$$

$$\hat{\theta}_t = \left(\sum_{s=1}^t A_s A_s^\top + \lambda I_d \right)^{-1} \sum_{s=1}^t A_s X_s$$

Deviation Inequality for Least Squares Estimates

Theorem (Confidence Ellipsoid)

For all $\delta > 0$,

$$\mathbb{P}(\forall t, \|\hat{\theta}_t - \theta^*\|_{V_t} \leq \beta_t) \geq 1 - \delta$$

where

$$V_t = \sum_{s=1}^t A_s A_s^\top + \lambda_t I_d$$

and

$$\beta_t = \sqrt{\lambda} S + \sigma \sqrt{2 \log\left(\frac{1}{\delta}\right) + d \log\left(1 + \frac{L^2 t}{d\lambda}\right)}$$

[Abbasi-Yadkori et al., 2001]

Lin-UCB Algorithm

Algorithm 1: Lin-UCB

Initialization: $b = 0_{\mathbb{R}^d}$, $V = \lambda I_d$, $\hat{\theta} = 0_{\mathbb{R}^d}$

for $t \geq 1$ **do**

 Receive \mathcal{A}_t

 Compute $\beta_{t-1} = \sqrt{\lambda} S + \sigma \sqrt{2 \log\left(\frac{1}{\delta}\right) + d \log\left(1 + \frac{L^2(t-1)}{\lambda d}\right)}$

for $a \in \mathcal{A}_t$ **do**

 Compute $\text{UCB}(a) = a^\top \hat{\theta} + \beta_{t-1} \sqrt{a^\top V^{-1} a} = \max_{\{\theta: \|\hat{\theta} - \theta\|_V \leq \beta_{t-1}\}} a^\top \theta$

$A_t = \operatorname{argmax}_a (\text{UCB}(a))$

Play action A_t **and receive reward** X_t

Updating: $V = V + A_t A_t^\top$, $b = b + X_t A_t$, $\hat{\theta} = V^{-1} b$

