Reinforcement Learning
TD 3 - Bandits

Fabien Pesquerel fabien.pesquerel@inria.fr
Philippe Preux philippe.preux@inria.fr

November 8, 2021

1 Problem - Structured and unstructured bandits

We study the classical multi-armed bandit problem specified by a set of real-valued (Gaussian or Bernoulli) distributions $(\nu_a)_{a\in\mathcal{A}}$ with means $(\mu_a)_{a\in\mathcal{A}} \in \mathbb{R}^{\mathcal{A}}$ and unitary variances, where $\mathcal{A}$ is a finite set of arms. We denote $\mu_* = \max\{\mu_a | a \in \mathcal{A}\}$.

At each time $t \geq 1$, an agent must choose an arm $a_t \in \mathcal{A}$, based only on the past. A reward X_t is drawn from the chosen distribution μ_{a_t} and observed by the agent. The goal of the agent is to maximize the expected sum of rewards received over time, or equivalently to minimize regret with respect to the strategy constantly receiving the highest mean reward.

1.1 Bandit Environement

The *Lai and Robbins lower bound* tells us that the regret is **asymptotically** no smaller than

$$\left(\sum_{a\in\mathcal{A}} \frac{\Delta_a}{\mathrm{KL}(\mu_a, \mu_*)}\right) \log(T),$$

where $\mathrm{KL}(\mu_a, \mu_*)$ is the KL-divergence between the Gaussian distribution of mean μ_a and the Gaussian distribution of mean μ_* (unitary variances). The constant in front of the $\log(T)$ may be called the **complexity** of the bandit problem.

For all arm $a \in \mathcal{A}$, for all time step $t \geq 1$, the empirical mean of arm a at time step t is

$$\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} \mathbb{1}_{\{a_s=a\}} X_s, \text{ if } N_a(t) > 0, \ 0 \text{ otherwise},$$

where $N_a(t) = \sum_{s=1}^{t} \mathbb{1}_{\{a_s=a\}}$ is the number of pulls of arm a at time t. We write $\hat{\mu}_*(t) = \max_{a\in\mathcal{A}} \hat{\mu}_a(t)$.

IMED strategy - Honda and Takemura (2011)

For an arm $a \in \mathcal{A}$ and a time step $t \geq 1$, the IMED index is defined as follows:

$$I_a(t) = N_a(t)\,\mathrm{KL}(\hat{\mu}_a(t), \hat{\mu}_*(t)) + \log(N_a(t)) \ .$$

This quantity can be seen as a transportation cost for "moving" a sub-optimal arm to an optimal one, plus exploration terms (the logarithms of the numbers of pulls). When an optimal arm is considered, the transportation cost is null and it remains only the exploration part. Note that, as stated in Honda and Takemura (2011), $I_a(t)$ is an index in a weak sense since it cannot be determined only by samples from the arm a but also uses empirical means of current optimal arms.

IMED is the strategy that consists in pulling an arm with minimal index at each time step:

$$a_{t+1} \in \operatorname*{argmin}_{a\in\mathcal{A}} I_a(t) \, .$$

KL-UCB strategy - Lai, Robbins, Cappé, Garivier, Maillard, Stoltz et al.

For an arm $a \in \mathcal{A}$ and a time step $t \geq 1$, the IMED index is defined as follows:

$$I_a(t) = \sup\left\{\mu \middle| \ N_a(t)\,\mathrm{KL}(\hat{\mu}_a(t), \mu) \leq \log t + 3\log\max\left(1, \log t\right)\right\} \ .$$

This quantity can be seen as an upper confidence bound on the empirical mean $\hat{\mu}_a(t)$.
KL-UCB is the strategy that consists in pulling an arm with minimal index at each time step:

$$a_{t+1} \in \operatorname*{argmax}_{a \in \mathcal{A}} I_a(t)\,.$$

1.2 Tools

Question 1 Derive the formula of the Kullback-Leibler divergences between two Gaussian distributions.

Question 2 Derive the formula of the Kullback-Leibler divergences between two Bernoulli distributions.

Question 3 Using `utils.py` and `bandits.py`, instantiate two bandits problems:

- One with 7 Gaussian distributions with means $\{0.1, 0.3, 0.4, 0.75, 0.8, 0.9, 0.95\}$
- One with 7 Bernoulli distributions with means $\{0.1, 0.3, 0.4, 0.75, 0.8, 0.9, 0.95\}$

Experiments Along this practical session, we will add algorithms to a common experimental test bed in order to compare the different sampling schemes. We will use the two bandit instances. We will use a time horizon of 1000. An experiment consists in using a sampling scheme on a bandit instance until the time horizon has been reached. During an experiment, one can (should) gather the cumulative regret. For each algorithm (insofar as they have been implemented) and bandit instance, you will run 200 experiments, plot the mean regret as well as a standard deviation tube around the mean regret curve. This will be called a **bandit experiment**.

1.3 Baselines

Question 4 Implement a round-robin sampling scheme algorithm.

Question 5 Implement the Explore Then Commit algorithm (commit at time step 300).

Question 6 Run a bandit experiment.

Question 7 Comment the plots.

1.4 Upper Confidence Bound algorithm

Question 8 Implement the **UCB** algorithm (use $\alpha = 2$) .

Question 9 Run a bandit experiment.

Question 10 Comment the plots.

1.5 IMED and KL-UCB

Question 11 Implement the **IMED** algorithm.

Question 12 Implement the **KL-UCB** algorithm.

Question 13 Run a bandit experiment.

Question 14 Comment the plots.

1.6 Thompson sampling

Question 15 Implement the Thompson sampling algorithm for Gaussian distributions.

Question 16 Implement the Thompson sampling algorithm for Bernoulli distributions.

Question 17 Run a bandit experiment.

Question 18 Comment the plots.

1.7 Lower bound

Question 19 Add the regret lower bound to a final bandit experiment. What can you say?

1.8 Unimodal Bandit

We assume that $\mathcal{A} = \{0, \dots, A-1\}, A \geq 1$, and $\mu : \begin{cases} \mathcal{A} & \to \mathbb{R} \\ a & \mapsto \mu_a \end{cases}$ is unimodal. That is, there exists $a_* \in \mathcal{A}$ such that $\mu_{[\![0,a_*]\!]}$ is increasing and $\mu_{[\![a_*,A]\!]}$ is decreasing. It is further assumed that for each arm a, ν_a is a Gaussian distribution $\mathcal{N}(\mu_a, 1)$, where $\mu_a \in \mathbb{R}$ is the mean of the distribution ν_a. We denote the structured set of Gaussian unimodal bandit distributions by

$$\mathcal{D}_{\text{unimodal}} = \left\{ \nu = (\nu_a)_{a \in \mathcal{A}} : \ \forall a \in \mathcal{A}, \ \nu_a \sim \mathcal{N}(\mu_a, 1) \text{ with } \mu_a \in \mathbb{R} \text{ and } \mu \text{ is unimodal} \right\}.$$

On a Gaussian unimodal bandit instance, the Lai and Robbins lower bound tells us that the regret is **asymptotically** no smaller than

$$\left(\sum_{a \in \mathcal{V}_{a_*}} \frac{\Delta_a}{\mathrm{KL}(\mu_a, \mu_*)} \right) \log(T),$$

where $\mathcal{V}_{a_*} = \{a_* - 1, a_* + 1\} \cap \mathcal{A}$ and $\mathrm{KL}(\mu_a, \mu_*)$ is the KL-divergence between the Gaussian distribution of mean μ_a and the Gaussian distribution of mean μ_* (variances equal to 1). The constant in front of the $\log(T)$ may be called the **unimodal complexity** of the bandit problem.

1.8.1 Computing regret lower bound for a unimodal bandit instance

Question Write a function that computes the complexity of a unimodal Gaussian bandit instance.

Question Write a function that generates at random a unimodal Gaussian bandit instance.

Question On a unimodal Gaussian bandit instance ν of your choice, add the theoretical lower bound $t \mapsto C_{\text{unimodal}}(\nu) \log(t)$ where $C_{\text{unimodal}}(\nu)$ is the unimodal complexity of a unimodal bandit problem to the regret curve of IMED. Add a plot with experiments of your choice using the previous algorithms.

1.8.2 IMED for Unimodal Bandit

For an arm $a \in \mathcal{A}$ and a time step $t \geq 1$, the IMED4UB index is defined as follows:

$$I_a(t) = N_a(t) \, \mathrm{KL}(\hat{\mu}_a(t), \hat{\mu}_*(t)) + \log(N_a(t)) \ .$$

IMED4UB is the strategy that consists in pulling an arm in the neighbourhood of the current optimal arm (also called leader arm, the one with the largest empirical mean) with minimal index at each time step:

$$a_{t+1} = \arg \min_{a \in \mathcal{V}_{\hat{a}_*(t)} \cup a_*} I_a(t) \, .$$

Question Write a class that provides an IMED type strategy for unimodal bandit inspired from the regret lower bound for unimodal structure.
Experiment like it was done in **part 1**. You can comment and add experiments of your choice.

1.9 Lipschitz Bandit

We assume that $\mathcal{A} = \{0, \dots, A-1\}, A \geq 1$, and $\mu : \begin{cases} \mathcal{A} & \to \mathbb{R} \\ a & \mapsto \mu_a \end{cases}$ is k-Lipschitz, where is k is assumed to be known. That is, for all $a, a' \in \mathcal{A}$, $|\mu_a - \mu_{a'}| \leq k \times |a - a'|$. It is further assumed that for each arm a, ν_a is a Gaussian distribution $\mathcal{N}(\mu_a, 1)$, where $\mu_a \in \mathbb{R}$ is the mean of the distribution ν_a. We denote the structured set of Gaussian k-Lipschitz bandit distributions by

$$\mathcal{D}_{k\text{-Lip}} = \left\{ \nu = (\nu_a)_{a \in \mathcal{A}} : \ \forall a \in \mathcal{A}, \ \nu_a \sim \mathcal{N}(\mu_a, 1) \text{ with } \mu_a \in \mathbb{R} \text{ and } \mu \text{ is } k\text{-Lipschitz} \right\}.$$

On a Gaussian k-Lipschitz bandit instance, the lower bounds tell us that the numbers of pulls satisfy **asymptotically** the following inequalities

$$\forall a \in \mathcal{A}, \quad \sum_{a' \in \mathcal{V}_a} \mathrm{KL}(\mu_{a'}, \, \mu_* - k|a - a'|) \ N_{a'}(T) \geq \log(T) \, ,$$

where $\mathcal{V}_a = \left\{ a' \in \mathcal{A} \colon \mu_{a'} < \mu_* - k|a - a'| \right\}$ and $\mathrm{KL}(\mu, \mu')$ is the KL-divergence between the Gaussian distribution of mean μ and the Gaussian distribution of mean μ' (unitary variances). The constant $C_{k\text{-Lip}}(\nu)$ resulting from the following linear programming problem may be called the **Lipschitz complexity** of the bandit problem:

$$C_{k\text{-Lip}}(\nu) = \min \left\{ \sum_{a \in \mathcal{A}} (\mu_* - \mu_a) \, n_a : \ n \in \mathbb{R}_+^{\mathcal{A}} \text{ s.t. } \forall a \in \mathcal{A}, \sum_{a' \in \mathcal{V}_a} \mathrm{KL}(\mu_{a'}, \, \mu_* - k|a - a'|) \ n_{a'} \geq 1 \right\}.$$

1.9.1 Computing regret lower bound for a k-Lipschitz bandit instance

Question Write a function that computes the complexity of a k-Lipschitz bandit instance.

Question Write a function that generates at random a Lipschitz Gaussian bandit instance.

Question On a k-Lipschitz Gaussian bandit instance ν of your choice, add the theoretical lower bound $t \mapsto C_{k-\mathrm{Lip}}(\nu) \log(t)$ where $C_{k\text{-Lip}}(\nu)$ is the lipschitz complexity of a lipschitz bandit problem to the regret curve of IMED.

1.9.2 IMED for Lipschitz Bandit

For an arm $a \in \mathcal{A}$ and a time step $t \geq 1$, the IMED4LB index is defined as foollows:

$$I_a(t) = \sum_{a' \in \hat{\mathcal{V}}_a(t)} N_{a'}(t) \, \mathrm{KL}(\hat{\mu}_{a'}(t), \hat{\mu}_*(t) - k|a - a'|) + \log(N_{a'}(t)) \ ,$$

where $\hat{\mathcal{V}}_a(t) = \left\{ a' \in \mathcal{A} \colon \hat{\mu}_{a'}(t) \leq \hat{\mu}_*(t) - k|a - a'| \right\}$.

IMED4LB is the strategy that consists in pulling an arm with minimal index at each time step:

$$a_{t+1} = \arg \min_{a \in \mathcal{V}_{\hat{a}_*(t)}} I_a(t) \, .$$

Question Write a class that provides an IMED type strategy for Lipschitz bandit inspired from the lower bounds on the number of pulls for Lipschitz structure.
Experiment like it was done in **part 1**. You can comment and add experiments of your choice.