Reinforcement Learning
TD 2 - Structure and Linearity

Fabien Pesquerel
fabien.pesquerel@inria.fr
Patrick Saux
patrick.saux@inria.fr
Odalric-Ambrym Maillard
odalric.maillard@inria.fr

January 6, 2020

1 Questions from the audience

You can write below the questions that were asked during the session and see later if you can remember/re-derive the answers.

2 Problem - Structured and unstructured bandits

Depending on your choice, please refer to the jupyter notebook or the python files.

We study the classical multi-armed bandit problem specified by a set of real-valued Gaussian 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.

2.1 Bandit Environement

On a Gaussian bandit instance, 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.

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

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:

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

2.1.1 Solving the exercise: How to?

Please try to refer to the jupyter notebook `TD2.ipynb` or the html version of the notebook. Otherwise, you should undersand and execute `TD2.py`. This code depends on the forban module.

1. Read, understand and execute the `how_to.ipynb` notebook.
2. Read, understand and execute the beginning of `TD2.ipynb`.

2.2 Upper Confidence Bound algorithms

Question 1 Using the `SequentiAlg` class, implement the **UCB** algorithm.

Question 2 Using the `Experiment` class, run experiments on (at least) two bandit problems of your choice.

Question 3 Comment the plots (suboptimality gaps, number of arms...).

2.3 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.

2.3.1 Computing regret lower bound for a unimodal bandit instance

Question 4 Write a function that computes the complexity of a unimodal Gaussian bandit instance.

Question 5 Write a function that generates at random a unimodal Gaussian bandit instance.

Question 6 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.

2.3.2 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 7 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.

2.4 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:

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

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

Question 8 Write a function that computes the complexity of a k-Lipschitz bandit instance.

Question 9 Write a function that generates at random a Lipschitz Gaussian bandit instance.

Question 10 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.

2.4.2 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 11 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.

2.5 Linear Bandit

Let us consider the discretisation $\mathcal{A} = \{0, \dots, A-1\}, A \geq 1$ of the space $\mathcal{X} = \left\{ x_a = \frac{a}{A} \ , a \in \mathcal{A} \right\} \subset [0, 1]$. Let us consider the parameter space $\Theta = \mathcal{B}(O, 1) \subset \mathbb{R}^d$, $d = 2p + 1$, $p \geq 1$, and the related function space $\mathcal{F}_\Theta = \left\{ f_\theta : x \in \mathcal{X} \mapsto \theta^{\mathrm{T}} \phi(x) \ , \theta \in \Theta \right\}$, where $\forall x \in \mathcal{X}, \phi(x) = (1, \cos(2\pi x), \sin(2\pi x), \dots, \cos(2\pi p x), \sin(2\pi p x))$. It is further assumed that for all parameter $\theta \in$

Θ, for all arm $a \in \mathcal{A}$, $\nu_a(\theta)$ is a Gaussian distribution $\mathcal{N}(\mu_a(\theta), 1)$, where $\mu_a(\theta) = f_\theta(x_a)$ is the mean of the distribution $\nu_a(\theta)$. We denote the structured set of Linear bandit distributions by

$$\mathcal{D}_\Theta = \left\{ \nu(\theta) = (\nu_a(\theta))_{a \in \mathcal{A}} \ , \theta \in \Theta \right\}.$$

Question 12 Write a function that generates at random a Linear Gaussian bandit instance.

2.5.1 The LinearUCB algorithm

Question 13 Implement the linear UCB algorithm: LinearUCB($\lambda = 1$, $\delta = 0.05$, $\sigma^2 = 1$).

2.5.2 The Linear IMED algorithm

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

$$I_a(t) = N_a^{\text{eff}}(t) \, \frac{\left(\max_{x \in \mathcal{X}} \hat{f}_{t,\lambda}(x) - \hat{f}_{t,\lambda}(x_a) \right)^2}{2} + \log\left(N_a^{\text{eff}}(t) \right) \, ,$$

where $\hat{f}_{t,\lambda}$ is the current estimate of f and $N_a^{\text{eff}}(t) = \left(\|\phi(x_a)\|^2_{G_t^{-1}} \right)^{-1}$, the effective number of pull of arms a. G_t is the current Gramm matrix (used for the regression).

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

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

Question 14 Write a class that provides an IMED type strategy for linear bandit inspired from the lower bounds on the number of pulls for Linear structure.

Question 15 Compare LinearUCB and LinearIMED.