Multi-Armed Bandits with Applications

Alexandra Carpentier
Uni Potsdam

October 30, 2023

Introduction

Sequential learning for an agent :

- Taking decisions in real time and in an uncertain environment...
- ...that influence the observations of the agent and its future actions.

Simplest sequential learning setting : bandit setting.

In this talk: Study of several bandit scenarii in different contexts.

Introduction

Sequential learning for an agent :

- Taking decisions in real time and in an uncertain environment...
- ...that influence the observations of the agent and its future actions.

Simplest sequential learning setting : bandit setting.
In this talk: Study of several bandit scenarii in different contexts.

Bandit setting

Simple mathematical framework for modeling some sequential decision making problems.

Play between many slot machines and maximise your earnings!

Outline

Cumulative regret

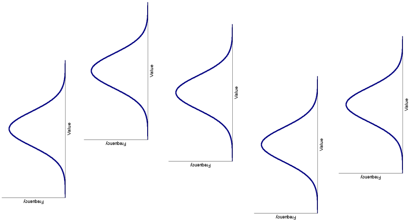
Best arm identification

Thresholding bandit problem

Bandit setting : the cumulative objective

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Cappé et al. (2013)], [Munos (2014)], etc.

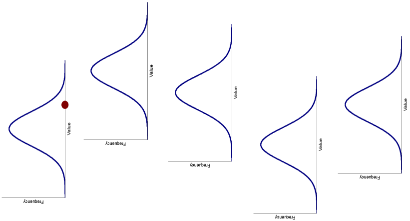
- K arms mechanisms
- Limited sampling resources T
- At each time t, choose k_t and collect X_t generated by mechanism k_t
- Objective : maximize $L_T = \sum_{t=1}^{T} X_t$



Bandit setting : the cumulative objective

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Cappé et al. (2013)], [Munos (2014)], etc.

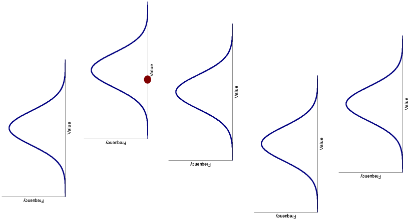
- K arms mechanisms
- Limited sampling resources T
- At each time t, choose k_t and collect X_t generated by mechanism k_t
- Objective : maximize $L_T = \sum_{t=1}^{T} X_t$



Bandit setting : the cumulative objective

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Cappé et al. (2013)], [Munos (2014)], etc.

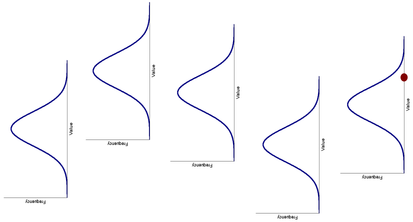
- K arms mechanisms
- Limited sampling resources T
- At each time t, choose k_t and collect X_t generated by mechanism k_t
- Objective : maximize $L_T = \sum_{t=1}^{T} X_t$



Bandit setting : the cumulative objective

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Cappé et al. (2013)], [Munos (2014)], etc.

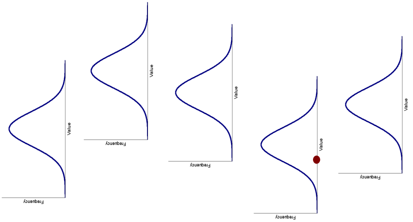
- K arms mechanisms
- Limited sampling resources T
- At each time t, choose k_t and collect X_t generated by mechanism k_t
- Objective : maximize $L_T = \sum_{t=1}^{T} X_t$



Bandit setting : the cumulative objective

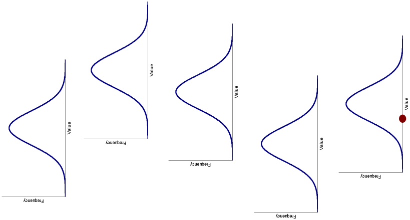
Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Cappé et al. (2013)], [Munos (2014)], etc.

- K arms mechanisms
- Limited sampling resources T
- At each time t, choose k_t and collect X_t generated by mechanism k_t
- Objective : maximize $L_T = \sum_{t=1}^T X_t$



Bandit setting : the cumulative objective

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Cappé et al. (2013)], [Munos (2014)], etc.

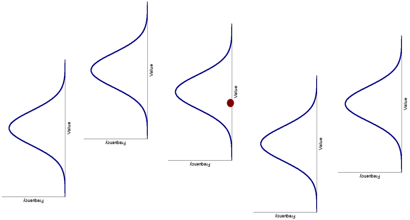


- K arms mechanisms
- Limited sampling resources T
- At each time t, choose k_t and collect X_t generated by mechanism k_t
- Objective : maximize $L_T = \sum_{t=1}^{T} X_t$

Bandit setting : the cumulative objective

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Cappé et al. (2013)], [Munos (2014)], etc.

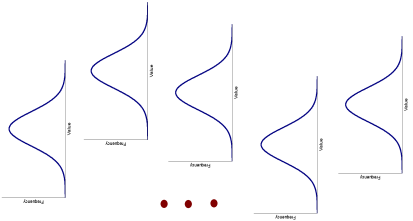
- K arms mechanisms
- Limited sampling resources T
- At each time t, choose k_t and collect X_t generated by mechanism k_t
- Objective : maximize $L_T = \sum_{t=1}^{T} X_t$



Bandit setting : the cumulative objective

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Cappé et al. (2013)], [Munos (2014)], etc.

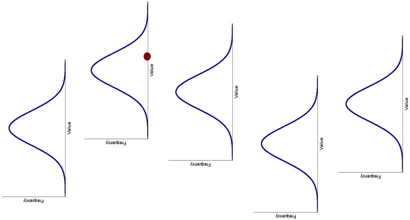
- K arms mechanisms
- Limited sampling resources T
- At each time t, choose k_t and collect X_t generated by mechanism k_t
- Objective : maximize $L_T = \sum_{t=1}^{T} X_t$



Bandit setting : the cumulative objective

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Cappé et al. (2013)], [Munos (2014)], etc.

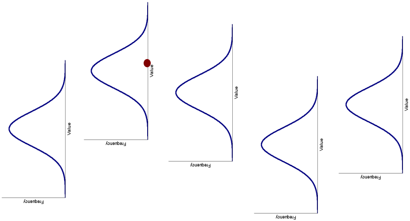
- K arms mechanisms
- Limited sampling resources T
- At each time t, choose k_t and collect X_t generated by mechanism k_t
- Objective : maximize $L_T = \sum_{t=1}^{T} X_t$



Bandit setting : the cumulative objective

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Cappé et al. (2013)], [Munos (2014)], etc.

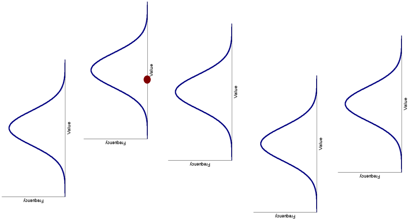
- K arms mechanisms
- Limited sampling resources T
- At each time t, choose k_t and collect X_t generated by mechanism k_t
- Objective : maximize $L_T = \sum_{t=1}^{T} X_t$



Bandit setting : the cumulative objective

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Cappé et al. (2013)], [Munos (2014)], etc.

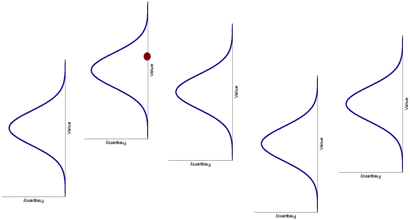
- K arms mechanisms
- Limited sampling resources T
- At each time t, choose k_t and collect X_t generated by mechanism k_t
- Objective : maximize $L_T = \sum_{t=1}^{T} X_t$



Bandit setting : the cumulative objective

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Cappé et al. (2013)], [Munos (2014)], etc.

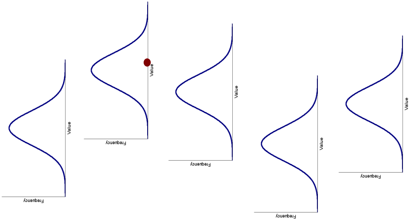
- K arms mechanisms
- Limited sampling resources T
- At each time t, choose k_t and collect X_t generated by mechanism k_t
- Objective : maximize $L_T = \sum_{t=1}^{T} X_t$



Bandit setting : the cumulative objective

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Cappé et al. (2013)], [Munos (2014)], etc.

- K arms mechanisms
- Limited sampling resources T
- At each time t, choose k_t and collect X_t generated by mechanism k_t
- Objective : maximize $L_T = \sum_{t=1}^{T} X_t$



Bandit setting : the cumulative objective

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Cappé et al. (2013)], [Munos (2014)], etc.

- K arms mechanisms
- Limited sampling resources T
- At each time t, choose k_t and collect X_t generated by mechanism k_t
- Objective : maximize $L_T = \sum_{t=1}^{T} X_t$

Applications : Historically, medical trials. Now rather used in recommender systems.

Bandit setting : the cumulative objective

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Cappé et al. (2013)], [Munos (2014)], etc.

- K arms mechanisms
- Limited sampling resources T
- At each time t, choose k_t and collect X_t generated by mechanism k_t
- Objective : maximize $L_T = \sum_{t=1}^{T} X_t$

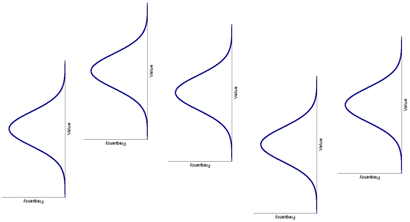
Problem : Need to learn the characteristic of the distribution while trying to allocate the samples to the best distribution! *Exploration/exploitation dilemma.*

Bandit setting : the cumulative objective

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Cappé et al. (2013)], [Munos (2014)], etc.

- K arms mechanisms
- Limited sampling resources T
- At each time t, choose k_t and collect X_t generated by mechanism k_t
- Objective : maximize $L_T = \sum_{t=1}^{T} X_t$

Bandit vocabulary:



Expected regret and notations: stochastic setting

Stohastic setting: arm mechanisms are K distributions $(\nu_k)_k$ that produce independent samples. Let us write

- μ_k for the mean of distribution k
- $\Delta_k = \max_i \mu_i - \mu_k$ gap of arm k
- $k^* \in \arg\max_i \mu_i$ optimal arm
- $T_{k,t}$ for the number of times distribution k has been sampled at time t
- $\hat{\mu}_{k,t} = \frac{1}{T_{k,t}} \sum_t X_t \mathbf{1}\{k_t = k\}$ for the empirical mean of distribution k at time t

Finite budget objective : Minimize the expected regret at time n

$$\mathbb{E}R_T = T \max_{k \le K} \mu_k - \mathbb{E} \sum_{t=1}^{T} X_t.$$

Classical assumption : The ν_k are supported on $[0,1]$.

Proposed strategies

Many strategies have been proposed as e.g.

- Thompson sampling [Thompson, 1933]
- Gittins index [Gittins, 1979]
- **Optimism in face of uncertainty** [Auer et al., 2002]

Optimism in face of uncertainty

In doubt, take the option that *could* be the best.

Algorithm 1 : UCB strategy (Auer et al., 2002)

Initialisation : Sample each distribution once.
For $t = 1...T$
  Set $k_t \in \arg\max[\hat{\mu}_{k,t} + 2\sqrt{\frac{\log(T)}{T_{k,t}}}]$
  Sample $X_t \sim \nu_{k_t}$
  Actualise $\hat{\mu}_{k,t}$ and $T_{k,t}$
EndFor

Exploration and exploitation!

Regret bounds for this algorithm

Theorem (Auer et al. , 2002)

The UCB strategy satisfies

$$\mathbb{E}R_T \leq 16 \sum_k \frac{\log(T)}{\Delta_k},$$

for $\Delta_k = \max_i \mu_i - \mu_k$ *and*

$$\mathbb{E}R_T \leq 32\sqrt{TK\log(T)},$$

Almost matching lower bounds - there exists an algorithm that reaches $\sqrt{TK}$, see [Bubeck et al, 2010].

Proof idea

High proba. event on the emp. means: Hoeffding + union bound gives

$$\mathbb{P}\left(\xi = \left\{\forall k, t : |\hat{\mu}_{k,t} - \mu_k| \leq 2\sqrt{\frac{\log(T)}{T_{k,t}}}\right\}\right) \geq 1 - 1/T^2.$$

Bounds on the number of arm pulls on ξ: At the last time t that a sub-optimal arm is pulled

$$\mu_k + 4\sqrt{\frac{\log(T)}{T_{k,n} - 1}} = \mu_k + 4\sqrt{\frac{\log(T)}{T_{k,t}}} \geq B_{k,t} \geq B_{k^*,t} \geq \mu_{k^*},$$

which implies $T_{k,n} \leq 1 + 16\frac{\log T}{\Delta_k^2}$.

Bound on the regret: Thus

$$R_T = \sum_k \Delta_k \mathbb{E} T_{k,n} \leq \sum_k \Delta_k (1 + 16\frac{\log T}{\Delta_k^2}) + 1/T.$$

Proof idea

High proba. event on the emp. means: Hoeffding + union bound gives

$$\mathbb{P}\left(\xi = \left\{\forall k, t : |\hat{\mu}_{k,t} - \mu_k| \leq 2\sqrt{\frac{\log(T)}{T_{k,t}}}\right\}\right) \geq 1 - 1/T^2.$$

Bounds on the number of arm pulls on ξ: At the last time t that a sub-optimal arm is pulled

$$\mu_k + 4\sqrt{\frac{\log(T)}{T_{k,n} - 1}} = \mu_k + 4\sqrt{\frac{\log(T)}{T_{k,t}}} \geq B_{k,t} \geq B_{k^*,t} \geq \mu_{k^*},$$

which implies $T_{k,n} \leq 1 + 16\frac{\log T}{\Delta_k^2}$.

Bound on the regret: Thus

$$R_T = \sum_k \Delta_k \mathbb{E} T_{k,n} \leq \sum_k \Delta_k (1 + 16\frac{\log T}{\Delta_k^2}) + 1/T.$$

Proof idea

High proba. event on the emp. means: Hoeffding + union bound gives

$$\mathbb{P}\left(\xi = \left\{\forall k, t : |\hat{\mu}_{k,t} - \mu_k| \le 2\sqrt{\frac{\log(T)}{T_{k,t}}}\right\}\right) \ge 1 - 1/T^2.$$

Bounds on the number of arm pulls on ξ: At the last time t that a sub-optimal arm is pulled

$$\mu_k + 4\sqrt{\frac{\log(T)}{T_{k,n}-1}} = \mu_k + 4\sqrt{\frac{\log(T)}{T_{k,t}}} \ge B_{k,t} \ge B_{k^*,t} \ge \mu_{k^*},$$

which implies $T_{k,n} \le 1 + 16\frac{\log T}{\Delta_k^2}$.

Bound on the regret: Thus

$$R_T = \sum_k \Delta_k \mathbb{E} T_{k,n} \le \sum_k \Delta_k (1 + 16\frac{\log T}{\Delta_k^2}) + 1/T.$$

Summary

Summary cumulative regret:

Regret R_T	prob. dep.	prob. indep.
	$\sum_k \frac{\log T}{\Delta_k}$	$\sqrt{TK}$

Expected regret and notations: adversarial setting

Adversarial setting: arm mechanisms generate K arbitrary sequence $(X_{k,t})$ in $[0,1]$.

Finite budget objective : Minimize the expected regret at time n

$$\bar{R}_T = \max_{k \leq K} \sum_{t \leq n} X_{k,t} - \mathbb{E}\Big[\sum_{t=1}^{T} X_t\Big].$$

Theorem (Auer et al. , 2002)

The EXP3 strategy satisfies

$$\bar{R}_T \leq 50\sqrt{TK\log(K)}.$$

Heavy use of randomisation to trick the environment (in case it is hostile).

Summary

Summary cumulative regret:

Regret R_T	prob. dep.	prob. indep.
	$\sum_k \frac{\log T}{\Delta_k}$	$\sqrt{TK}$

Outline

Cumulative regret

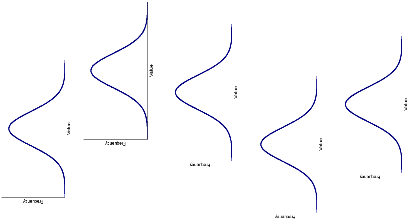
Best arm identification

Thresholding bandit problem

Sequential learning

Resource allocation in face of uncertainty : See [Robbins (1952)], [Gittins (1979)], [Whittle, 1988], [Cappé et al. (2013)], [Munos (2014)], etc.

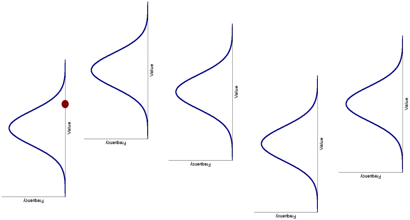
- Distributions $(\nu_k)_{k \leq K}$ supported on $[0, 1]$ and with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Aim : maximise a function of the collected data $(X_1, \ldots, X_T)$



Sequential learning

Resource allocation in face of uncertainty : See [Robbins (1952)], [Gittins (1979)], [Whittle, 1988], [Cappé et al. (2013)], [Munos (2014)], etc.

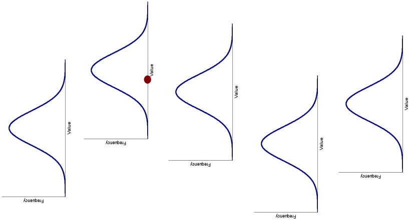
- Distributions $(\nu_k)_{k\leq K}$ supported on $[0,1]$ and with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Aim : maximise a function of the collected data $(X_1, \ldots, X_T)$



Sequential learning

Resource allocation in face of uncertainty : See [Robbins (1952)], [Gittins (1979)], [Whittle, 1988], [Cappé et al. (2013)], [Munos (2014)], etc.

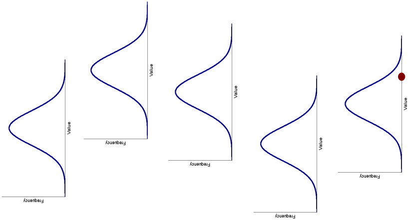
- Distributions $(\nu_k)_{k\leq K}$ supported on $[0,1]$ and with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Aim : maximise a function of the collected data $(X_1, \ldots, X_T)$



Sequential learning

Resource allocation in face of uncertainty : See [Robbins (1952)], [Gittins (1979)], [Whittle, 1988], [Cappé et al. (2013)], [Munos (2014)], etc.

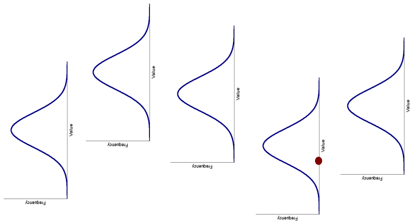
- Distributions $(\nu_k)_{k \leq K}$ supported on $[0,1]$ and with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Aim : maximise a function of the collected data $(X_1, \ldots, X_T)$



Sequential learning

Resource allocation in face of uncertainty : See [Robbins (1952)], [Gittins (1979)], [Whittle, 1988], [Cappé et al. (2013)], [Munos (2014)], etc.

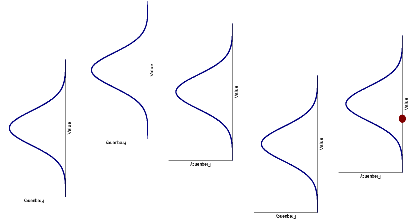
- Distributions $(\nu_k)_{k\leq K}$ supported on $[0,1]$ and with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Aim : maximise a function of the collected data $(X_1, \ldots, X_T)$



Sequential learning

Resource allocation in face of uncertainty : See [Robbins (1952)], [Gittins (1979)], [Whittle, 1988], [Cappé et al. (2013)], [Munos (2014)], etc.

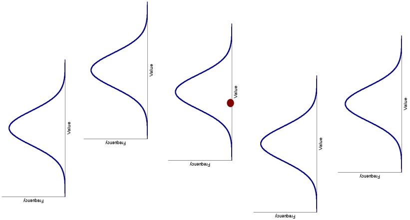
- Distributions $(\nu_k)_{k\leq K}$ supported on $[0,1]$ and with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Aim : maximise a function of the collected data $(X_1,\ldots,X_T)$



Sequential learning

Resource allocation in face of uncertainty : See [Robbins (1952)], [Gittins (1979)], [Whittle, 1988], [Cappé et al. (2013)], [Munos (2014)], etc.

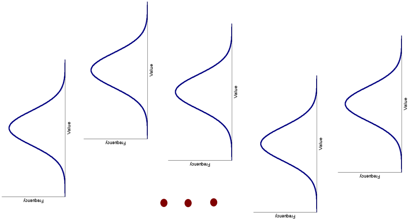
- Distributions $(\nu_k)_{k\leq K}$ supported on $[0,1]$ and with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Aim : maximise a function of the collected data $(X_1, \dots, X_T)$



Sequential learning

Resource allocation in face of uncertainty : See [Robbins (1952)], [Gittins (1979)], [Whittle, 1988], [Cappé et al. (2013)], [Munos (2014)], etc.

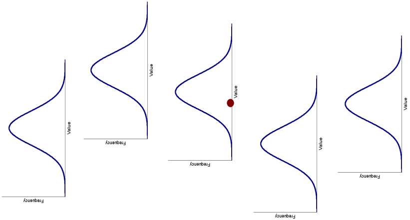
- Distributions $(\nu_k)_{k \leq K}$ supported on $[0,1]$ and with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Aim : maximise a function of the collected data $(X_1, \ldots, X_T)$



Sequential learning

Resource allocation in face of uncertainty : See [Robbins (1952)], [Gittins (1979)], [Whittle, 1988], [Cappé et al. (2013)], [Munos (2014)], etc.

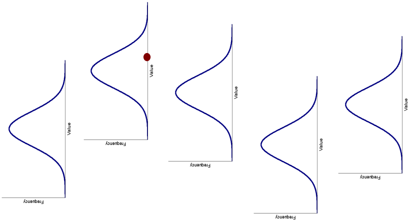
- Distributions $(\nu_k)_{k \leq K}$ supported on $[0, 1]$ and with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Aim : maximise a function of the collected data $(X_1, \ldots, X_T)$



Sequential learning

Resource allocation in face of uncertainty : See [Robbins (1952)], [Gittins (1979)], [Whittle, 1988], [Cappé et al. (2013)], [Munos (2014)], etc.

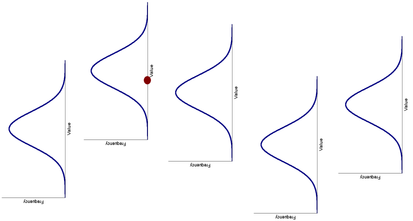
- Distributions $(\nu_k)_{k \leq K}$ supported on $[0, 1]$ and with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Aim : maximise a function of the collected data $(X_1, \ldots, X_T)$



Sequential learning

Resource allocation in face of uncertainty : See [Robbins (1952)], [Gittins (1979)], [Whittle, 1988], [Cappé et al. (2013)], [Munos (2014)], etc.

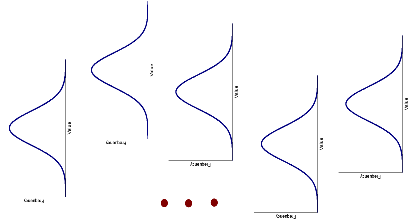
- Distributions $(\nu_k)_{k\leq K}$ supported on $[0,1]$ and with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Aim : maximise a function of the collected data $(X_1, \ldots, X_T)$



Sequential learning

Resource allocation in face of uncertainty : See [Robbins (1952)], [Gittins (1979)], [Whittle, 1988], [Cappé et al. (2013)], [Munos (2014)], etc.

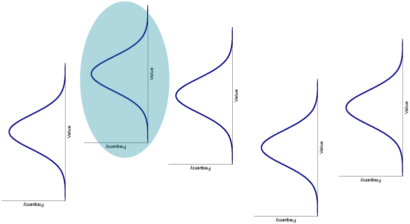
- Distributions $(\nu_k)_{k \leq K}$ supported on $[0, 1]$ and with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Aim : maximise a function of the collected data $(X_1, \ldots, X_T)$



Sequential learning

Resource allocation in face of uncertainty : See [Robbins (1952)], [Gittins (1979)], [Whittle, 1988], [Cappé et al. (2013)], [Munos (2014)], etc.

- Distributions $(\nu_k)_{k \leq K}$ supported on $[0, 1]$ and with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Aim : maximise a function of the collected data $(X_1, \ldots, X_T)$



Sequential learning

Resource allocation in face of uncertainty : See [Robbins (1952)], [Gittins (1979)], [Whittle, 1988], [Cappé et al. (2013)], [Munos (2014)], etc.

- Distributions $(\nu_k)_{k \leq K}$ supported on $[0, 1]$ and with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Aim : maximise a function of the collected data $(X_1, \ldots, X_T)$

Info. theoretic question

Given that we can collect T data as we want, how well can we achieve our objective?

Sequential learning

Resource allocation in face of uncertainty : See [Robbins (1952)], [Gittins (1979)], [Whittle, 1988], [Cappé et al. (2013)], [Munos (2014)], etc.

- Distributions $(\nu_k)_{k\leq K}$ supported on $[0,1]$ and with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Aim : maximise a function of the collected data $(X_1,\ldots,X_T)$

Info. theoretic question

Given that we can collect T data as we want, how well can we achieve our objective?

Answear

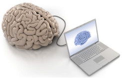
Characterize the best possible algorithmic performance given the sequential collection of T data.

Sequential learning

Resource allocation in face of uncertainty : See [Robbins (1952)], [Gittins (1979)], [Whittle, 1988], [Cappé et al. (2013)], [Munos (2014)], etc.

- Distributions $(\nu_k)_{k \leq K}$ supported on $[0,1]$ and with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Aim : maximise a function of the collected data $(X_1, \ldots, X_T)$

Best arm identif.: output $\hat{k}$ and find $k^* = \arg\max \mu_k$.



Sequential learning

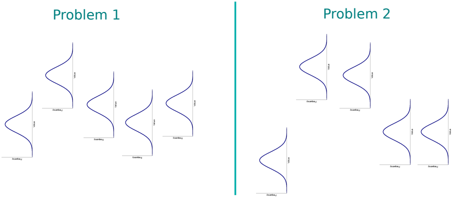
Resource allocation in face of uncertainty : See [Robbins (1952)], [Gittins (1979)], [Whittle, 1988], [Cappé et al. (2013)], [Munos (2014)], etc.

- Distributions $(\nu_k)_{k\leq K}$ supported on $[0,1]$ and with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Aim : maximise a function of the collected data $(X_1, \dots, X_T)$

Best arm identif.: output $\hat{k}$ and find $k^* = \arg\max \mu_k$.

Question

Smallest possible $\mathbb{P}(\hat{k} \neq k^*)$ achieved by an algorithm given that we can collect T data?



Sequential learning

Resource allocation in face of uncertainty : See [Robbins (1952)], [Gittins (1979)], [Whittle, 1988], [Cappé et al. (2013)], [Munos (2014)], etc.

- Distributions $(\nu_k)_{k \leq K}$ supported on $[0, 1]$ and with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Aim : maximise a function of the collected data $(X_1, \ldots, X_T)$

Best arm identif.: output $\hat{k}$ and find $k^* = \arg\max \mu_k$.

Smallest error: will depend on the distance between the distribution's means.

Model complexity :

$$H := \sum_{k \neq k^*} \frac{1}{(\mu_{k^*} - \mu_k)^2}.$$

Known H

[Audibert and Bubeck, 2010]'s strategy : based on an UCB

$$k_t = \arg\max_k [\hat{\mu}_{k,t} + \sqrt{\frac{aT}{T_{k,t}}}], \quad \begin{cases} \hat{\mu}_{k,t} & \text{empirical mean} \\ T_{k,t} & \text{nb. of collected samples.} \end{cases}$$

At time T, recommend

$$\hat{k} \in \arg\max_k \hat{\mu}_{k,T}.$$

Theorem (Audibert and Bubeck, 2010, Kaufmann et. al, 2015, C. and Locatelli, 2016)

If

$$1/a = \mathbf{H} := \sum_{\mathbf{k} \neq \mathbf{k}^*} \frac{\mathbf{1}}{(\mu_{\mathbf{k}^*} - \mu_{\mathbf{k}})^{\mathbf{2}}},$$

then

$$\mathbb{P}(\hat{k} \neq k^*) \leq \square \exp(-\square TH).$$

For any H, any strategy, there exists a problem such that

$$\mathbb{P}(\hat{k} \neq k^*) \geq \square \exp(-\square TH).$$

Unknown H

[Audibert and Bubeck, 2010]'s "agnostic" strategy : divide the budget T in $\log(K)$ and run with $\log(K)$ well-chosen parameters a. Then aggregate samples.

Theorem (Audibert and Bubeck, 2010)

For this "agnostic" strategy

$$\mathbb{P}(\hat{k} \neq k^*) \leq \square \exp(-\square \frac{TH}{\log(K)}).$$

Theorem (C. and Locatelli, 2016)

For any strategy there exists a problem such that

$$\mathbb{P}(\hat{k} \neq k^*) \geq \square \exp(-\square \frac{TH}{\log(K)}).$$

Summary

Summary cumulative regret:

Regret R_T	prob. dep.	prob. indep.
	$\sum_k \frac{\log T}{\Delta_k}$	$\sqrt{TK}$

Summary best arm identification:

Status of H	$\mathbb{P}(\hat{k} \neq k^*)$	$r_T = \mu^* - \mu_{\hat{k}}$
Known	$\square \exp(-\square TH)$	$\sqrt{T/K}$
Unknown	$\square \exp(-\square TH/\log(K))$	$\sqrt{K/T}$

Outline

Cumulative regret

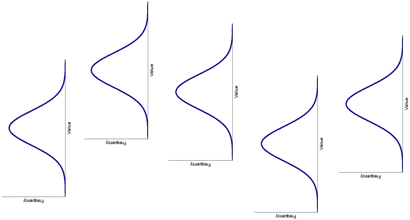
Best arm identification

Thresholding bandit problem

Sequential learning

Resource optimisation in face of uncertainty : See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Whittle (1988)], [Cappé et al. (2013)], [Munos (2014)], etc.

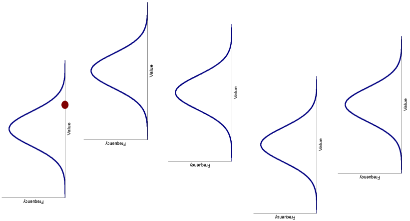
- Distributions $(\nu_k)_{k \leq K}$ with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Objective to fulfil



Sequential learning

Resource optimisation in face of uncertainty : See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Whittle (1988)], [Cappé et al. (2013)], [Munos (2014)], etc.

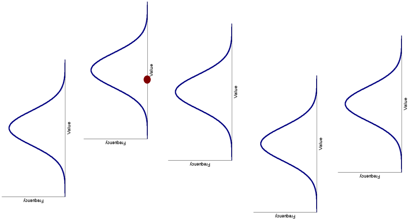
- Distributions $(\nu_k)_{k \leq K}$ with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Objective to fulfil



Sequential learning

Resource optimisation in face of uncertainty : See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Whittle (1988)], [Cappé et al. (2013)], [Munos (2014)], etc.

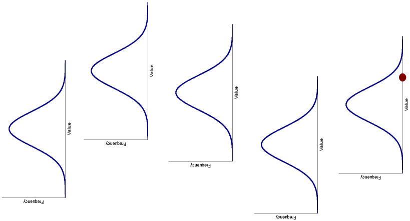
- Distributions $(\nu_k)_{k \leq K}$ with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Objective to fulfil



Sequential learning

Resource optimisation in face of uncertainty : See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Whittle (1988)], [Cappé et al. (2013)], [Munos (2014)], etc.

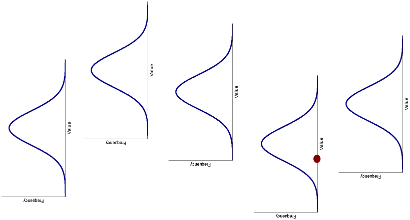
- Distributions $(\nu_k)_{k \leq K}$ with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Objective to fulfil



Sequential learning

Resource optimisation in face of uncertainty : See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Whittle (1988)], [Cappé et al. (2013)], [Munos (2014)], etc.

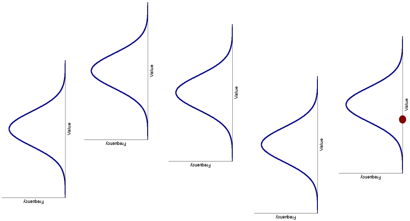
- Distributions $(\nu_k)_{k \leq K}$ with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Objective to fulfil



Sequential learning

Resource optimisation in face of uncertainty : See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Whittle (1988)], [Cappé et al. (2013)], [Munos (2014)], etc.

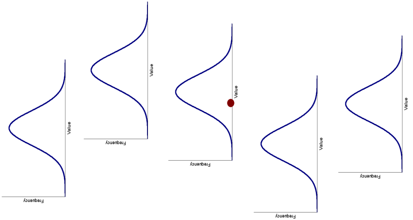
- Distributions $(\nu_k)_{k \le K}$ with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Objective to fulfil



Sequential learning

Resource optimisation in face of uncertainty : See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Whittle (1988)], [Cappé et al. (2013)], [Munos (2014)], etc.

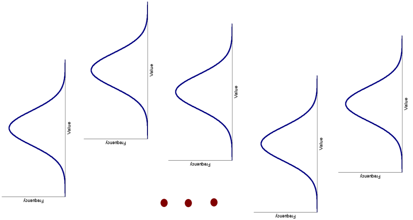
- Distributions $(\nu_k)_{k \leq K}$ with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Objective to fulfil



Sequential learning

Resource optimisation in face of uncertainty : See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Whittle (1988)], [Cappé et al. (2013)], [Munos (2014)], etc.

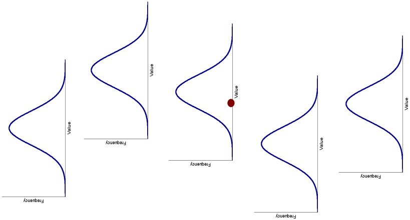
- Distributions $(\nu_k)_{k \leq K}$ with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Objective to fulfil



Sequential learning

Resource optimisation in face of uncertainty : See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Whittle (1988)], [Cappé et al. (2013)], [Munos (2014)], etc.

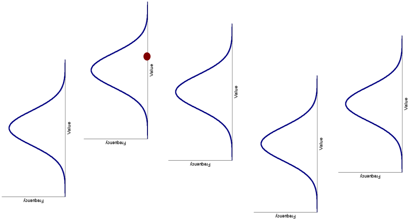
- Distributions $(\nu_k)_{k \leq K}$ with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Objective to fulfil



Sequential learning

Resource optimisation in face of uncertainty : See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Whittle (1988)], [Cappé et al. (2013)], [Munos (2014)], etc.

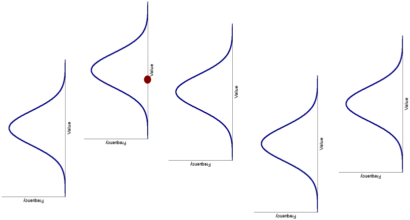
- Distributions $(\nu_k)_{k \le K}$ with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Objective to fulfil



Sequential learning

Resource optimisation in face of uncertainty : See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Whittle (1988)], [Cappé et al. (2013)], [Munos (2014)], etc.

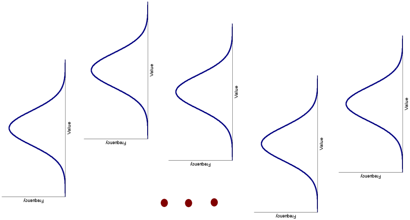
- Distributions $(\nu_k)_{k \leq K}$ with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Objective to fulfil



Sequential learning

Resource optimisation in face of uncertainty : See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Whittle (1988)], [Cappé et al. (2013)], [Munos (2014)], etc.

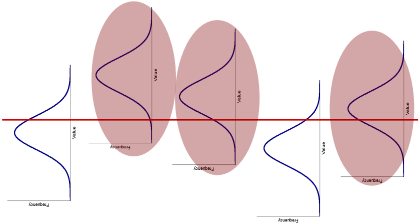
- Distributions $(\nu_k)_{k \leq K}$ with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Objective to fulfil



Sequential learning

Resource optimisation in face of uncertainty : See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Whittle (1988)], [Cappé et al. (2013)], [Munos (2014)], etc.

- Distributions $(\nu_k)_{k \leq K}$ with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Objective to fulfil



Sequential learning

Resource optimisation in face of uncertainty : See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], [Whittle (1988)], [Cappé et al. (2013)], [Munos (2014)], etc.

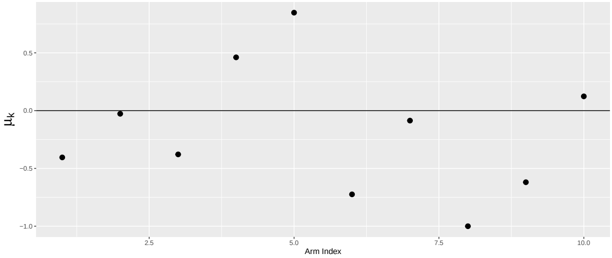
- Distributions $(\nu_k)_{k \leq K}$ with *unknown* means μ_k
- Limited sampling resources T
- At each time t, choose k_t and collect $X_t \sim \nu_{k_t}$
- Objective to fulfil

This is the *thresholding bandit problem*, i.e. given a threshold τ, and writing μ_k for the mean of distribution k, we aim at predicting

$$Q = (\mathrm{sign}(\mu_k - \tau))_k \, .$$

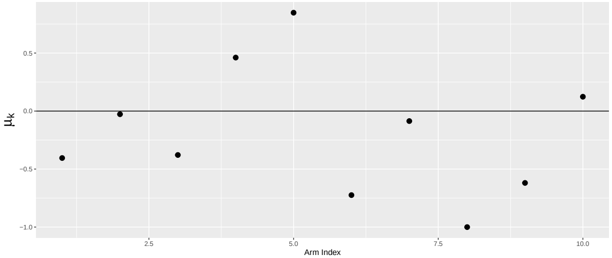
Problem setting: K arms, budget T, threshold $\tau = 0$

- Each arm $k \in [K]$ corresponds to a distribution $\mathcal{N}(\mu_k, 1)$ with mean $\mu_k \in [-1, 1]$ - and we set $\tau = 0$.
- At each round $t < T$ the learner pulls an arm $k_t \in [K]$ and observes a sample $X_t \sim \mathcal{N}(\mu_{k_t}, 1)$.
- Upon exhaustion of the budget the learner is required to output a prediction $\hat{Q} \in \{-1, 1\}^K$ of $Q = \text{sign}(\mu_k)$.



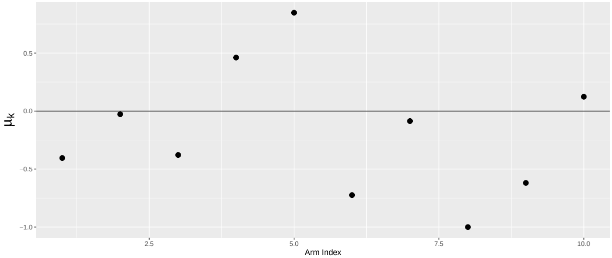
Problem setting: K arms, budget T, threshold $\tau = 0$

- Each arm $k \in [K]$ corresponds to a distribution $\mathcal{N}(\mu_k, 1)$ with mean $\mu_k \in [-1, 1]$ - and we set $\tau = 0$.
- At each round $t < T$ the learner pulls an arm $k_t \in [K]$ and observes a sample $X_t \sim \mathcal{N}(\mu_{k_t}, 1)$.
- Upon exhaustion of the budget the learner is required to output a prediction $\hat{Q} \in \{-1, 1\}^K$ of $Q = \mathrm{sign}(\mu_k)$.



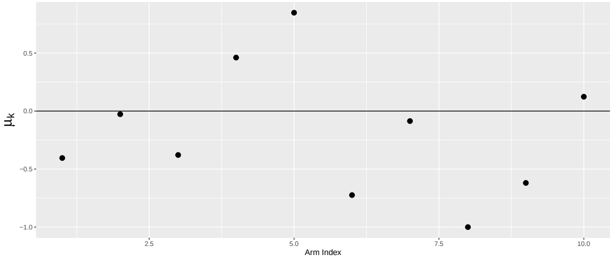
Problem setting: K arms, budget T, threshold $\tau = 0$

- Each arm $k \in [K]$ corresponds to a distribution $\mathcal{N}(\mu_k, 1)$ with mean $\mu_k \in [-1, 1]$ - and we set $\tau = 0$.
- At each round $t < T$ the learner pulls an arm $k_t \in [K]$ and observes a sample $X_t \sim \mathcal{N}(\mu_{k_t}, 1)$.
- Upon exhaustion of the budget the learner is required to output a prediction $\hat{Q} \in \{-1, 1\}^K$ of $Q = \text{sign}(\mu_k)$.



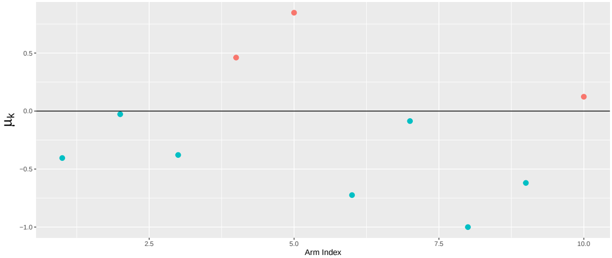
Problem setting: K arms, budget T, threshold $\tau = 0$

- Each arm $k \in [K]$ corresponds to a distribution $\mathcal{N}(\mu_k, 1)$ with mean $\mu_k \in [-1, 1]$ - and we set $\tau = 0$.
- At each round $t < T$ the learner pulls an arm $k_t \in [K]$ and observes a sample $X_t \sim \mathcal{N}(\mu_{k_t}, 1)$.
- Upon exhaustion of the budget the learner is required to output a prediction $\hat{Q} \in \{-1, 1\}^K$ of $Q = \text{sign}(\mu_k)$.

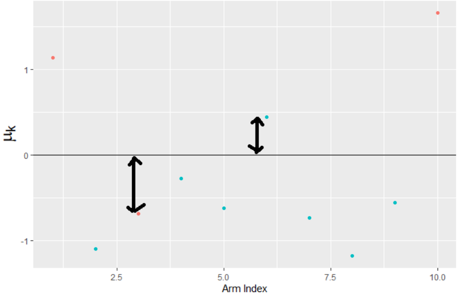


Problem setting: K arms, budget T, threshold $\tau = 0$

- Each arm $k \in [K]$ corresponds to a distribution $\mathcal{N}(\mu_k, 1)$ with mean $\mu_k \in [-1, 1]$ - and we set $\tau = 0$.
- At each round $t < T$ the learner pulls an arm $k_t \in [K]$ and observes a sample $X_t \sim \mathcal{N}(\mu_{k_t}, 1)$.
- Upon exhaustion of the budget the learner is required to output a prediction $\hat{Q} \in \{-1, 1\}^K$ of $Q = \text{sign}(\mu_k)$.



Regret



Two measures of regret:

- Probability of error:

$$e_T := \mathbb{P}(\hat{Q} \neq Q).$$

- Simple regret:

$$r_T := \mathbb{E} \max_{k:\hat{Q}[k] \neq Q[k]} |\mu_k|.$$

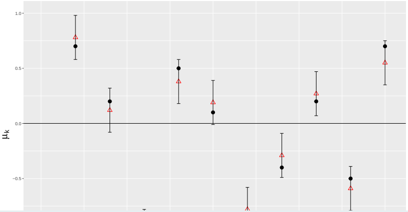
Problem independent results

Theorem (Cheshire et. al, 2020)

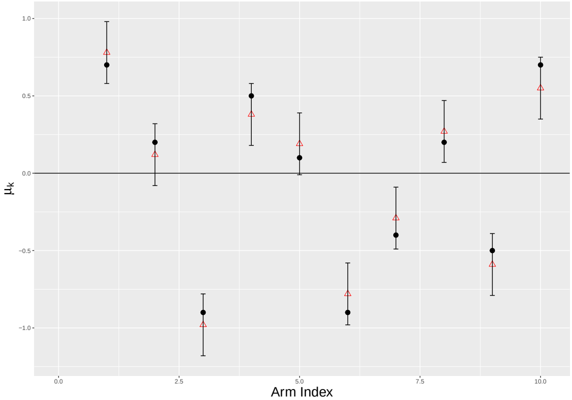
It holds that (uniform sampling reaches this)

$$\inf_{\text{algo}} \sup_{\text{problem}} r_T \asymp \sqrt{\frac{K \log(K)}{T}}.$$

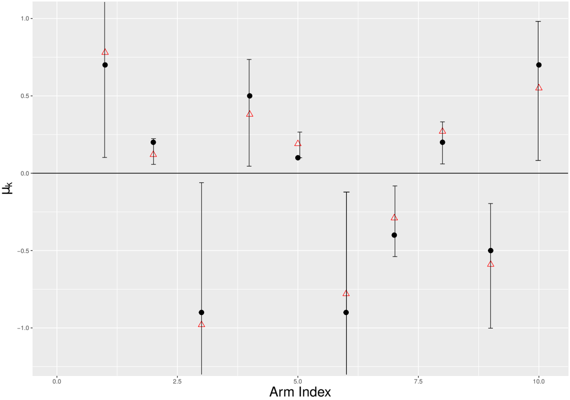
Upper bound trivial (uniform sampling), lower bound somewhat more tricky than in batch setting.



Unconstrained setting: problem dependent results



Unconstrained setting: problem dependent results



Unconstrained setting: problem dependent results

In what follows: write the gaps

$$\Delta_i = |\mu_i|,$$

and $\mathcal{M}_{\bar{\Delta}}$ the set of problems with gaps $\bar{\Delta}$.

Theorem (Locatelli et al., 2016)

For any vector of gaps $\bar{\Delta}$ it holds that

$$K \log(T) \exp(-\Box T/H) \gtrsim \inf_{\text{algo}} \sup_{\text{problem in } \mathcal{M}_{\bar{\Delta}}} e_T \gtrsim \exp(-\Box T/H),$$

where $H = \sum_i \bar{\Delta}_i^{-2}$.

Unconstrained setting: problem dependent results

In what follows: write the gaps

$$\Delta_i = |\mu_i|,$$

and $\mathcal{M}_{\bar{\Delta}}$ the set of problems with gaps $\bar{\Delta}$.

Theorem (Locatelli et al., 2016)

For any vector of gaps $\bar{\Delta}$ it holds that

$$K \log(T) \exp(-\square T/H) \gtrsim \inf_{\text{algo}} \sup_{\text{problem in } \mathcal{M}_{\bar{\Delta}}} e_T \gtrsim \exp(-\square T/H),$$

where $H = \sum_i \bar{\Delta}_i^{-2}$.

APT algorithm: sample at time t

$$k_t \in \arg\min_k T_{k,t} |\hat{\mu}_{k,t}|^2.$$

Conclusion

Theorem ((Locatelli et al, 2016), (Cheshire et al, 2020))

It holds that

$$\inf_{\text{algo}} \sup_{\text{problem}} r_T \approx \sqrt{\frac{K \log K}{T}},$$

and for $T \gtrsim \log K \vee \log\log T$ *and any* $\bar{\Delta}$

$$\inf_{\text{algo}} \sup_{\bar{\Delta}-\text{problem}} \log e_T \asymp -T/H.$$

Summary

Summary cumulative regret:

Regret R_T	prob. dep.	prob. indep.
	$\sum_k \frac{\log T}{\Delta_k}$	$\sqrt{TK}$

Summary thresholding bandit problem:

Regret R_T	prob. dep.	prob. indep.
	$\square \exp(-\square TH)$	$\sqrt{K \log K/T}$

Summary

Summary cumulative regret:

Regret R_T	prob. dep.	prob. indep.
	$\sum_k \frac{\log T}{\Delta_k}$	$\sqrt{TK}$

Summary best arm identification:

Status of H	$\mathbb{P}(\hat{k} \neq k^*)$	$r_T = \mu^* - \mu_{\hat{k}}$
Known	$\square \exp(-\square TH)$	$\sqrt{T/K}$
Unknown	$\square \exp(-\square TH/\log(K))$	$\sqrt{K/T}$

Summary thresholding bandit problem:

Regret R_T	prob. dep.	prob. indep.
	$\square \exp(-\square TH)$	$\sqrt{K \log K/T}$

Conclusion

In this talk:

- Three bandit problems: cumulative regret, best arm identification, thresholding bandit problem.
- Strategies: optimism in the face of uncertainty
- Slight change of assumptions between thresholding bandit and best arm identification: change in the optimal rate