IETR
CentraleSupélec

On-line learning for real-time dynamic spectrum access: from theory to practice

CRIStAL
Centre de Recherche en Informatique, Signal et Automatique de Lille

Stochastic Multi-Armed Bandit for Single User ...and Beyond

Christophe Moy and Emilie Kaufmann,

CentraleSupélec
CNRS

ICC Tutorial, May 25th, 2017

Multi-armed bandit setting

From a single device point of view:

channels: streams of rewards

Channel 1	$X_{1,1}$	$X_{1,2}$	$\dots$	$X_{1,t}$	$\dots$	$X_{1,T}$
Channel 2	$X_{2,1}$	$X_{2,2}$	$\dots$	$X_{2,t}$	$\dots$	$X_{2,T}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	
Channel K	$X_{K,1}$	$X_{K,2}$	$\dots$	$X_{K,t}$	$\dots$	$X_{K,T}$

Example:

- $X_{a,t} = 1$ or 0 if the communication is successful or unsuccessful on channel a at round t

At round t, the device:

- selects channel A_t
- receives the reward $X_t = X_{A_t,t}$

Multi-armed bandit setting

From a single device point of view:

***arms**: streams of rewards*

Arm 1	$X_{1,1}$	$X_{1,2}$	$\dots$	$X_{1,t}$	$\dots$	$X_{1,T}$
Arm 2	$X_{2,1}$	$X_{2,2}$	$\dots$	$X_{2,t}$	$\dots$	$X_{2,T}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	
Arm K	$X_{K,1}$	$X_{K,2}$	$\dots$	$X_{K,t}$	$\dots$	$X_{K,T}$

Example:

- $X_{a,t} = 1$ or 0 if the communication is successful or unsuccessful on channel a at round t

At round t, **an agent**:

- selects **arm** A_t
- receives the reward $X_t = X_{A_t,t}$

Outline

Stochastic bandit model and algorithms
- First algorithms
- UCB algorithms
- Bayesian algorithms

Beyond the stochastic MAB
- Relaxing the i.i.d. assumption
- Relaxing the stochastic assumption

Bandits for multiple devices

Stochastic multi-armed bandit model

A simple stochastic assumption:

$$\forall k = 1, \ldots, K, \quad (X_{k,t})_{t \in \mathbb{N}} \quad \text{is} \quad \text{i.i.d.} \quad \text{with a distribution } \nu_k$$

arm $\leftrightarrow$ (unknown) probability distribution

ν_1 ν_2 ν_3 ν_4 ν_5

At round t, an agent:

- chooses an arm A_t
- observes a reward $X_t = X_{A_t,t} \sim \nu_{A_t}$

The sampling strategy (or bandit algorithm) (A_t) is sequential:

$$A_{t+1} = F_t(A_1, X_1, \ldots, A_t, X_t).$$

Stochastic multi-armed bandit model

A simple stochastic assumption:

$$\forall k = 1, \dots, K, \quad (X_{k,t})_{t \in \mathbb{N}} \quad \text{is} \quad \text{i.i.d.} \quad \text{with a distribution } \nu_k$$

arm $\leftrightarrow$ (unknown) probability distribution

ν_1 ν_2 ν_3 ν_4 ν_5

At round t, an agent:

- chooses an arm A_t
- observes a reward $X_t = X_{A_t,t} \sim \nu_{A_t}$

Goal: find a strategy maximizing $\sum_{t=1}^{T} X_t$ (cumulated rewards)

Why bandits?

Historical motivation: clinical trials [Thompson 1933]

- arm $\leftrightarrow$ medical treatment

- Which treatment should be allocated to each patient based on the previously observed effects?

Why bandits?

Historical motivation: clinical trials **[Thompson 1933]**

- arm $\leftrightarrow$ medical treatment

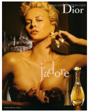


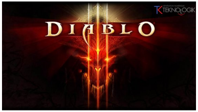



- Which treatment should be allocated to each patient based on the previously observed effects?

$$ **Motivation: online advertisement** **[2010 ...]**

- arm $\leftrightarrow$ add

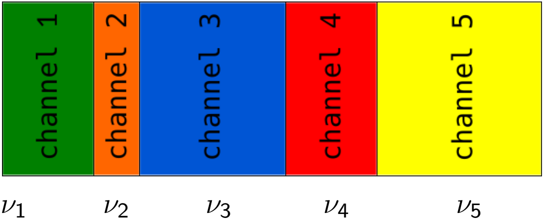




- Which add should be displayed to each visitor based on the previously observed clicks?

Stochastic bandits for cognitive radios

A frequency band:



What distributions for the arms?

- Bernoulli $\mathcal{B}(p_k)$ to model the channel availability

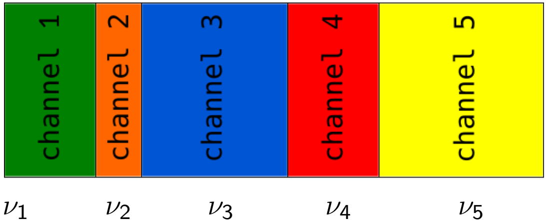
$$\mathbb{P}(X_{k,t} = 1) = p_k \quad \text{and} \quad \mathbb{P}(X_{k,t} = 0) = 1 - p_k$$

  p_k: mean availability of channel k (unknown!)

- Other possible distributions ν_k to model the quality of the communication, with mean p_k (e.g., ν_k is bounded)

Stochastic bandits for cognitive radios

A frequency band:



What distributions for the arms?

- Bernoulli $\mathcal{B}(p_k)$ to model the channel availability

$$\mathbb{P}(X_{k,t} = 1) = p_k \quad \text{and} \quad \mathbb{P}(X_{k,t} = 0) = 1 - p_k$$

  p_k: mean availability of channel k (unknown!)
- Other possible distributions ν_k to model the quality of the communication, with mean p_k (e.g., ν_k is bounded)

Objective

Goal: find a strategy maximizing

$$\mathbb{E}\left[\sum_{t=1}^{T} X_t\right].$$

Cognitive radios:

- maximize the (average) fraction of sucessfull transmissions
- maximize the (average) quality of the communications

Objective

Goal: find a strategy maximizing

$$\mathbb{E}\left[\sum_{t=1}^{T} X_t\right].$$

Cognitive radios:

- maximize the (average) fraction of sucessfull transmissions
- maximize the (average) quality of the communications

Oracle: always play the arm

$$k^* = \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}} \; p_k \quad \text{with mean} \quad p^* = \max_{k\in\{1,\dots,K\}} p_k.$$

Can we be *almost as good as the oracle*?

$$\mathbb{E}\left[\sum_{t=1}^{T} X_t\right] \simeq Tp^*?$$

Performance measure: the regret

Maximizing rewards $\leftrightarrow$ minimizing *regret*

$$R_T := Tp^* - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$$

$$= \sum_{k=1}^{K} (p^* - p_k)\mathbb{E}[T_k(T)],$$

$T_k(t)$: number of draws of arm k up to round t.

➜ Need for an Exploration/Exploitation tradeoff

Performance measure: the regret

Maximizing rewards $\leftrightarrow$ minimizing *regret*

$$R_T \quad := \quad Tp^* - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$$

We want the regret to *grow sub-linearly*:

$$\frac{R_T}{T} \underset{T\to\infty}{\longrightarrow} 0 \quad (\textit{consistency})$$

➜ what rate of regret can we expect?

A lower bound on the regret

Bernoulli bandit model, $\boldsymbol{p} = (p_1, \dots, p_K)$

$$R_T(\boldsymbol{p}) = \sum_{k=1}^{K} (p^* - p_k)\mathbb{E}_{\boldsymbol{p}}[T_k(T)]$$

When T grows, all the arms should be drawn infinitely many!

- [Lai & Robbins, 1985]: for any "uniformly good" strategy,

$$p_k < p^* \Rightarrow \liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{p}}[T_k(T)]}{\log T} \geq \frac{1}{\mathrm{d}(p_k, p^*)},$$

where

$$\begin{aligned} \mathrm{d}(p, p') &= \mathrm{KL}(\mathcal{B}(p), \mathcal{B}(p')) \\ &= p \log \frac{p}{p'} + (1-p) \log \frac{1-p}{1-p'}. \end{aligned}$$

➜ the regret is at least logarithmic

A lower bound on the regret

Bernoulli bandit model, $\boldsymbol{p} = (p_1, \ldots, p_K)$

$$R_T(\boldsymbol{p}) = \sum_{k=1}^{K} (p^* - p_k) \mathbb{E}_{\boldsymbol{p}}[T_k(T)]$$

When T grows, all the arms should be drawn infinitely many!

- [Lai & Robbins, 1985]: for any "uniformly good" strategy,

$$p_k < p^* \Rightarrow \liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{p}}[T_k(T)]}{\log T} \geq \frac{1}{\mathrm{d}(p_k, p^*)},$$

where

$$\begin{aligned} \mathrm{d}(p, p') &= \mathrm{KL}(\mathcal{B}(p), \mathcal{B}(p')) \\ &= p \log \frac{p}{p'} + (1-p) \log \frac{1-p}{1-p'}. \end{aligned}$$

➜ can we find asymptotically optimal algorithm, i.e. algorithms matching the lower bound?

Outline

Stochastic bandit model and algorithms

- First algorithms
- UCB algorithms
- Bayesian algorithms

Beyond the stochastic MAB

- Relaxing the i.i.d. assumption
- Relaxing the stochastic assumption

Bandits for multiple devices

Some (naive) strategies

- **Idea 1 :** Draw each arm T/K times

$\Rightarrow$ EXPLORATION

$$R(T) = \left(\frac{1}{K} \sum_{a=2}^{K} (p_1 - p_a) \right) T$$

Some (naive) strategies

- **Idea 1 :** Draw each arm T/K times

$\Rightarrow$ EXPLORATION

$$R(T) = \left(\frac{1}{K} \sum_{a=2}^{K} (p_1 - p_a) \right) T$$

- **Idea 2 :** Always trust the empirical best arm

$$A_{t+1} = \underset{k \in \{1,\dots,K\}}{\text{argmax}} \; \hat{X}_k(t)$$

where

$$\hat{X}_k(t) = \frac{\text{sum of the rewards observed from } k \text{ up to round } t}{\text{number of selections of } k \text{ up to round } t}$$

is an estimate of the unknown mean p_k.

$\Rightarrow$ EXPLOITATION

$$\mathbb{R}(T) \geq (1 - p_1) \times \mu_2 \times (p_1 - \mu_2) T$$

Some (naive) strategies

- **Idea 1 :** Draw each arm T/K times

$\Rightarrow$ EXPLORATION

$$R(T) = \left(\frac{1}{K} \sum_{a=2}^{K} (p_1 - p_a) \right) T$$

- **Idea 2 :** Always trust the empirical best arm

$$A_{t+1} = \underset{k \in \{1,\dots,K\}}{\text{argmax}} \; \hat{X}_k(t)$$

where

$$\hat{X}_k(t) = \frac{1}{T_k(t)} \sum_{s=1}^{t} X_s \mathbb{1}_{(A_s = k)}$$

is an estimate of the unknown mean p_k.

$\Rightarrow$ EXPLOITATION

$$\mathbb{R}(T) \geq (1 - p_1) \times \mu_2 \times (p_1 - \mu_2) T$$

A better idea: Explore-Then-Exploit

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

- draw each arm m times
- compute the empirical best arm $\hat{k} = \text{argmax}_k \hat{X}_k(Km)$
- keep playing this arm until round T

$$A_{t+1} = \hat{k} \text{ for } t \geq Km$$

$\Rightarrow$ EXPLORATION followed by EXPLOITATION

A better idea: Explore-Then-Exploit

Given $m \in \{1, \dots, T/K\}$,

- draw each arm m times
- compute the empirical best arm $\hat{k} = \text{argmax}_k \, \hat{X}_k(Km)$
- keep playing this arm until round T

$$A_{t+1} = \hat{k} \quad \text{for } t \geq Km$$

⇒ EXPLORATION followed by EXPLOITATION

Analysis: 2 arms, $p_1 > p_2$. $\Delta = p_1 - p_2$.

$$R_T = \Delta \times \mathbb{E}[T_2(T)]$$

$$\begin{aligned} T_2(T) &= m + (T - 2m)\mathbb{1}_{(\hat{k}=2)} \\ \mathbb{E}[T_2(T)] &\leq m + (T - 2m)\mathbb{P}\left(\hat{X}_1(2m) < \hat{X}_2(2m)\right) \\ &\leq m + T\exp\left(-\frac{m\Delta^2}{2}\right) \quad \textit{(Hoeffding's inequality)} \end{aligned}$$

A better idea: Explore-Then-Exploit

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

- draw each arm m times
- compute the empirical best arm $\hat{k} = \text{argmax}_k \hat{X}_k(Km)$
- keep playing this arm until round T

$$A_{t+1} = \hat{k} \text{ for } t \geq Km$$

$\Rightarrow$ EXPLORATION followed by EXPLOITATION

Analysis: 2 arms, $p_1 > p_2$. $\Delta = p_1 - p_2$.

$$R_T \leq \underbrace{\Delta m}_{\text{increases with } m} + \underbrace{\Delta T \exp\left(-\frac{m\Delta^2}{2}\right)}_{\text{decreases with } m}$$

A good choice: $m = \left\lfloor \frac{2}{\Delta^2} \log\left(\frac{T\Delta^2}{2}\right) \right\rfloor$

A better idea: Explore-Then-Exploit

Given $m \in \{1, \dots, T/K\}$,

- draw each arm m times
- compute the empirical best arm $\hat{k} = \text{argmax}_k \hat{X}_k(Km)$
- keep playing this arm until round T

$$A_{t+1} = \hat{k} \quad \text{for } t \geq Km$$

$\Rightarrow$ EXPLORATION followed by EXPLOITATION

Analysis: 2 arms, $p_1 > p_2$. $\Delta = p_1 - p_2$.

$$R_T \leq \frac{2}{\Delta}\left[\log\left(\frac{T\Delta^2}{2}\right) + 1\right]$$

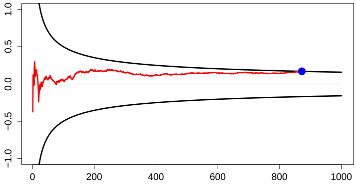
A good choice: $m = \left\lfloor \frac{2}{\Delta^2} \log\left(\frac{T\Delta^2}{2}\right) \right\rfloor$

➜ requires the knowledge of $\Delta = p_1 - p_2$!

Sequential Explore-Then-Exploit (2 arms)

- explore uniformly until the random time

$$\tau = \inf \left\{ t \in \mathbb{N} : |\hat{X}_1(t) - \hat{X}_2(t)| > \sqrt{\frac{4\log(T/t)}{t}} \right\}$$



- $\hat{k} = \mathrm{argmax}_k\ \hat{X}_k(\tau)$ and $(A_{t+1} = \hat{k})$ for $t \in \{\tau, \dots, T\}$

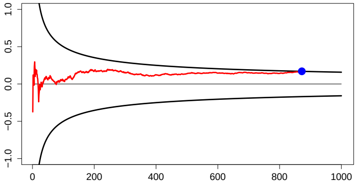
$$R_T \le \frac{2}{\Delta} \log(T) + C\sqrt{\log(T)}.$$

➜ same regret rate, without knowing Δ

Sequential Explore-Then-Exploit (2 arms)

- explore uniformly until the random time

$$\tau = \inf \left\{ t \in \mathbb{N} : |\hat{X}_1(t) - \hat{X}_2(t)| > \sqrt{\frac{4\log(T/t)}{t}} \right\}$$



- $\hat{k} = \mathrm{argmax}_k \ \hat{X}_k(\tau)$ and $(A_{t+1} = \hat{k})$ for $t \in \{\tau, \dots, T\}$

$$R_T \leq \frac{2}{\Delta} \log(T) + C\sqrt{\log(T)}.$$

➜ still requires the knowledge of T...

Outline

Stochastic bandit model and algorithms
First algorithms
UCB algorithms
Bayesian algorithms

Beyond the stochastic MAB
Relaxing the i.i.d. assumption
Relaxing the stochastic assumption

Bandits for multiple devices

The optimism principle

- For each arm k, assume we have a confidence interval on the unknown mean p_k :

$$\mathcal{I}_k(t) = [\mathrm{LCB}_k(t), \mathrm{UCB}_k(t)]$$

$$\mathrm{LCB} = \text{Lower Confidence Bound}$$
$$\mathrm{UCB} = \text{Upper Confidence Bound}$$

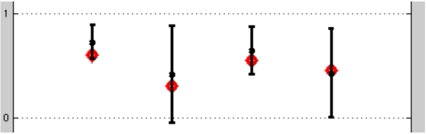


Figure: Confidence intervals on the means after t rounds

The optimism principle

- We apply the following principle:

"act as if the best possible model was the true model"

(optimism in face of uncertainty)

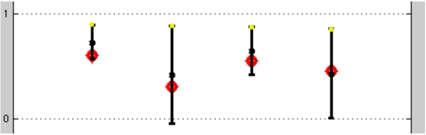


Figure: Confidence intervals on the means after t rounds

- Thus, one selects at time $t+1$ the arm

$$A_{t+1} = \underset{k=1,\dots,K}{\operatorname{argmax}} \, \mathrm{UCB}_k(t)$$

How to build the Confidence Intervals?

We need to build $\mathrm{U}_k(t)$ such that

$$\mathbb{P}\left(p_k \leq \mathrm{U}_k(t)\right) \gtrsim 1 - \frac{1}{t}.$$

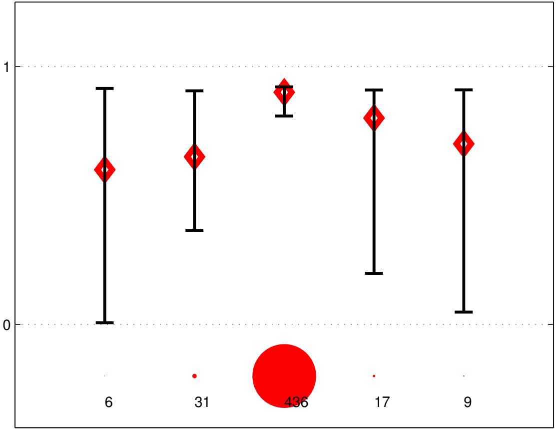
UCB1 [Auer et al. 02] chooses $A_{t+1} = \mathrm{argmax}_k\ \mathrm{U}_k(t)$ with

$$\mathrm{U}_k(t) = \underbrace{\hat{X}_k(t)}_{\text{exploitation term}} + \underbrace{\sqrt{\frac{2\log(t)}{T_k(t)}}}_{\text{exploration bonus}}.$$

(for distribution bounded in $[0, 1]$)

- tools: Hoeffding's inequality + a union bound
- a (simple !) *finite time* analysis

A UCB algorithm in practice



An improved analysis of UCB1

Define the index

$$\mathrm{U}_k(t) = \hat{X}_k(t) + \sqrt{\frac{\alpha \log(t)}{T_k(t)}}$$

Theorem [Bubeck '11],[Cappé et al.'13]

For $\alpha > 1/2$, the UCB algorithm using the above index satisfies

$$\mathbb{E}[T_k(T)] \leq \frac{\alpha}{(p_1 - p_a)^2} \log(T) + O(\sqrt{\log(T)}).$$

➜ "order-optimal" w.r.t. Lai and Robbins' lower bound

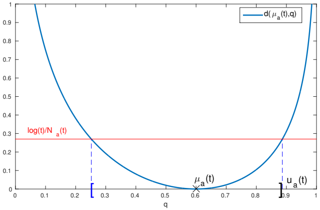
$\left[\text{Pinsker's inequality: } d(p_a, p_1) \geq 2(p_1 - p_a)^2\right]$

The kl-UCB algorithm

- A UCB-type algorithm: $A_{t+1} = \underset{k}{\arg\max}\ u_k(t)$
- ... associated to the right upper confidence bound:

$$u_k(t) = \max\left\{ q : d\left(\hat{X}_k(t), q\right) \leq \frac{\log(t)}{T_k(t)} \right\},$$

with $d(x,y) = \mathrm{KL}(\mathcal{B}(x), \mathcal{B}(y))$.



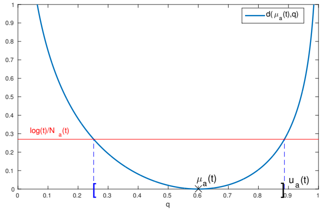
[Cappé et al. 13] : $\mathbb{E}_{\boldsymbol{\mu}}[T_k(T)] \leq \frac{1}{d(p_k, p^*)} \log T + O(\sqrt{\log(T)})$.

The kl-UCB algorithm

- A UCB-type algorithm: $A_{t+1} = \underset{k}{\text{argmax}}\ u_k(t)$
- ... associated to the right upper confidence bound:

$$u_k(t) = \max\left\{q : d\left(\hat{X}_k(t), q\right) \leq \frac{\log(t)}{T_k(t)}\right\},$$

with $d(x,y) = \mathrm{KL}(\mathcal{B}(x), \mathcal{B}(y))$.



- kl-UCB is asymptotically optimal for Bernoulli bandits!

Outline

Stochastic bandit model and algorithms

First algorithms

UCB algorithms

Bayesian algorithms

Beyond the stochastic MAB

Relaxing the i.i.d. assumption

Relaxing the stochastic assumption

Bandits for multiple devices

The Bayesian choice

Bernoulli bandit model $\boldsymbol{p} = (p_1, \ldots, p_K)$

- **frequentist view**: $p_1, \ldots, p_K$ are unknown parameters

➜ tools: estimators, confidence intervals

The Bayesian choice

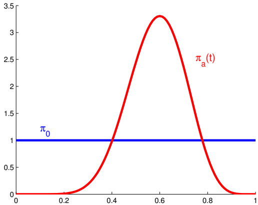
Bernoulli bandit model $\boldsymbol{p} = (p_1, \dots, p_K)$

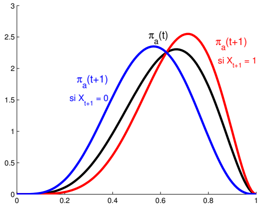
- **Bayesian view**: $p_1, \dots, p_K$ are random variables

  prior distribution : $p_a \sim \mathcal{U}([0,1])$

➜ tool: posterior distribution

$$\begin{aligned}\pi_k(t) &= \mathcal{L}\left(p_k | X_1, \dots, X_t\right) \\ &= \mathrm{Beta}(S_k(t)+1, T_k(t)-S_k(t)+1)\end{aligned}$$

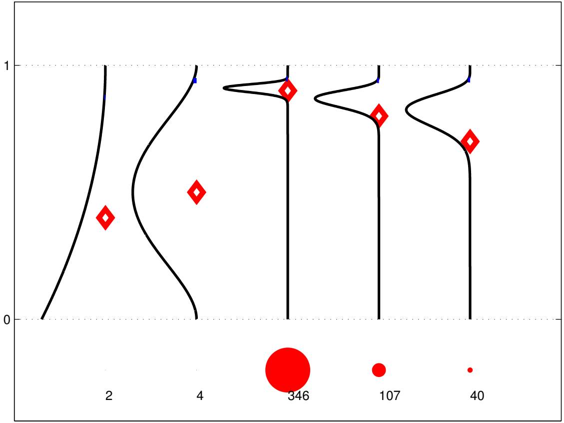




$S_k(t) = \sum_{s=1}^{t} X_s \mathbb{1}_{(A_s=k)}$ sum of the rewards from arm k

Bayesian algorithm

A Bayesian bandit algorithm exploits the posterior distributions of the means to decide which arm to select.



The Bayes-UCB algorithm

$\pi_k(t)$ the posterior distribution over p_k at the end of round t.

Bayes-UCB [K., Cappé, Garivier 2012] selects

$$A_{t+1} = \underset{k \in \{1,\dots,K\}}{\text{argmax}} \; Q\left(1 - \frac{1}{t}, \pi_k(t)\right)$$

where $Q(\alpha, \nu)$ is the quantile of order α of the distribution ν.

$$\mathbb{P}_{X \sim \nu}(X \leq Q(\alpha, \nu)) = \alpha.$$

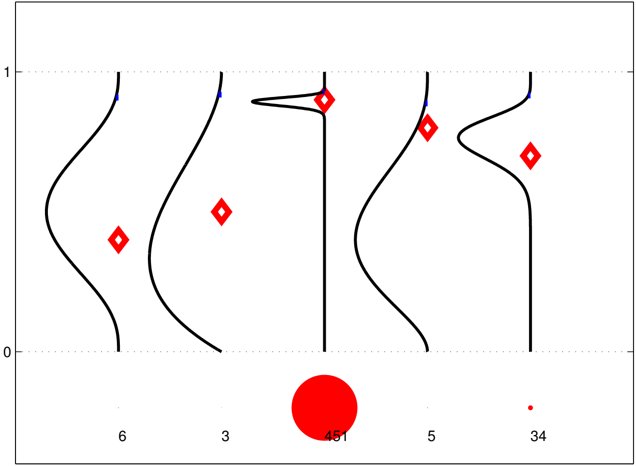
Properties:

- ➜ easy to implement (quantiles of Beta distributions)
- ➜ also asymptotically optimal for Bernoulli bandits!

$$q_k(t) = Q\left(1 - \frac{1}{t}, \pi_k(t)\right) \simeq u_k(t)$$

- ➜ efficient in practice and easy to generalize

Bayes-UCB in practice



Thompson Sampling

$$\begin{cases} \forall a \in \{1..K\}, \quad \theta_a(t) \sim \pi_a(t) \\ A_{t+1} = \underset{a=1\dots K}{\text{argmax}}\ \theta_a(t). \end{cases}$$

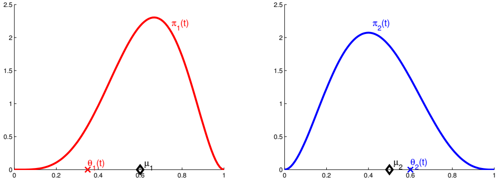
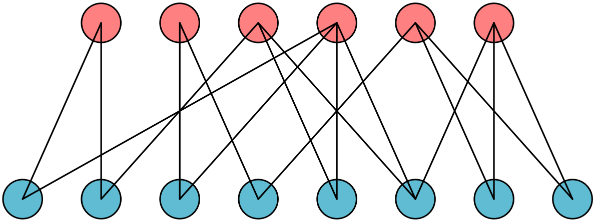


Figure: TS selects arm 2 as $\theta_2(t) \geq \theta_1(t)$

- ➜ the first bandit algorithm! [Thompson 1933]
- ➜ very efficient, beyond Bernoulli bandits
- ➜ matches the Lai and Robbins bound for Bernoulli bandits
  K., Korda and Munos, *Thompson Sampling: an Asymptotically Optimal Finite-Time Analysis*, ALT 2012

TS for *structured* stochastic bandits ?

- Arms are edges on a graph
- $\mathcal{M}$ is a set of possible configurations (subsets of edges)
- The agent chooses $m_t \in \mathcal{M}$ at time t and observe a realization of all arms in $\mathcal{M}$



Example: find a matching between users and channels
Lelarge et al., *Spectrum Bandit Optimization*, ITW 2013

Outline

Stochastic bandit model and algorithms

First algorithms

UCB algorithms

Bayesian algorithms

Beyond the stochastic MAB

Relaxing the i.i.d. assumption

Relaxing the stochastic assumption

Bandits for multiple devices

Outline

Stochastic bandit model and algorithms

- First algorithms
- UCB algorithms
- Bayesian algorithms

Beyond the stochastic MAB

- Relaxing the i.i.d. assumption
- Relaxing the stochastic assumption

Bandits for multiple devices

A restless bandit example

Restless Markov bandit

$$\text{for all } k, \quad (X_{k,t})_{t\in\mathbb{N}} \quad \text{is a Markov chain}$$

Cognitive radio:

- the behavior of primary users is evolving according to a Markovian dynamic
- simple model: state space $\{0, 1\}$ occupied/available

A restless bandit example

Restless Markov bandit

$$\text{for all } k, \quad (X_{k,t})_{t\in\mathbb{N}} \quad \text{is a Markov chain}$$

Cognitive radio:

- the behavior of primary users is evolving according to a Markovian dynamic
- simple model: state space $\{0, 1\}$ occupied/available

Idea: If arm k has *stationnary distribution* π_k, aim to always select the channel

$$\underset{k\in\{1,\dots,K\}}{\text{argmax}} \ \mathbb{E}_{X\sim\pi_k}[X]$$

➜ may be a bad idea...

Optimal strategies?

Example: a transition matrix on $\{0, 1\}$

$$P_\epsilon = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}$$

P_ϵ has invariant measure $\pi_\epsilon = [1/2, 1/2]$, with mean $1/2$.

2-armed bandit:

- arm 1 is a Markov chain with transition P_{ϵ_1}
- arm 2 is a Markov chain with transition P_{ϵ_2}

Strategies:

- static strategy playing a single arm $\rightarrow$ average reward $1/2$
- a much better strategy when ϵ_1 and ϵ_2 are small: switch arm when the current state is 0

➜ regret with respect to the best static action is no longer (always) the right notion

[Ryabko et al. 2014]

Some tools for restless bandits

- Bayesian approaches based on Whittle indices
  Liu and Zhao, Indexability of Restless Bandit Problems and Optimality of Whittle Index for Dynamic Multichannel Access. I.T., 2010
- Using reinforcement learning algorithms
  (restless Markov bandit = a Markov Decision Process)
  Ortner, Ryabko, Auer and Munos, Regret bounds for restless Markov bandits, TCS, 2014
- Can we modify the UCB approach?
  Liu et al, Learning in a Changing World: Restless Multiarmed Bandit With Unknown Dynamics. IEEE I.T., 2013

Experiments:

- plain UCB may still be robust on some Markovian arms

Outline

Stochastic bandit model and algorithms

First algorithms

UCB algorithms

Bayesian algorithms

Beyond the stochastic MAB

Relaxing the i.i.d. assumption

Relaxing the stochastic assumption

Bandits for multiple devices

A new bandit game

At round t

- the player chooses arm A_t
- simultaneously, an adversary chooses the vector of rewards

$$(x_{1,t}, \ldots, x_{K,t})$$

- the player receives the reward $x_t = x_{A_t,t}$

Goal: maximize rewards, or minimize regret

$$\mathrm{R}(T) = \max_a \mathbb{E}\left[\sum_{t=1}^{T} x_{a,t}\right] - \mathbb{E}\left[\sum_{t=1}^{T} x_t\right].$$

Exponential Weighted Forecaster

The full-information game: at round t

- the player chooses arm A_t
- simultaneously, an adversary chooses the vector of rewards

$$(x_{t,1}, \ldots, x_{t,K})$$

- the player receives the reward $x_t = x_{A_t,t}$
- and he observes the reward vector $(x_{t,1}, \ldots, x_{t,K})$

The EWF algorithm **[Littelstone, Warmuth 1994]**

With $\hat{p}_t$ the probability distribution

$$\hat{p}_t(k) \propto e^{\eta\left(\sum_{s=1}^{t-1} x_{k,s}\right)}$$

at round t, choose

$$A_t \sim \hat{p}_t$$

The EXP3 strategy

We don't have access to the $(x_{k,t})$ for all k...

$$\hat{x}_{k,t} = \frac{x_{k,t}}{\hat{p}_{k,t}} \mathbb{1}_{(A_t = k)}$$

satisfies $\mathbb{E}[\hat{x}_{k,t}] = x_{a,t}$.

The EXP3 algorithm

With $\hat{p}_t$ the probability distribution

$$\hat{p}_t(k) \propto e^{\eta\left(\sum_{s=1}^{t-1} \hat{x}_{k,s}\right)}$$

at round t, choose

$$A_t \sim \hat{p}_t$$

Auer, Cesa-Bianchi, Freund, Schapire, *The nonstochastic multiarmed bandit problem*, SIAM J. Comput., 2002

Theoretical results

The EXP3 strategy

With $\hat{p}_t$ the probability distribution

$$\hat{p}_t(k) \propto e^{\eta\left(\sum_{s=1}^{t-1} \hat{x}_{k,s}\right)}$$

at round t, choose

$$A_t \sim \hat{p}_t$$

Theorem [Bubeck and Cesa-Bianchi 12]
EXP3 with

$$\eta = \sqrt{\frac{\log(K)}{KT}}$$

satisfies

$$R(T) \leq \sqrt{2 \log K}\sqrt{KT}$$

Outline

Stochastic bandit model and algorithms
First algorithms
UCB algorithms
Bayesian algorithms

Beyond the stochastic MAB
Relaxing the i.i.d. assumption
Relaxing the stochastic assumption

Bandits for multiple devices

Multi-players bandits: setup

M players playing *the same* K-armed bandit $(M \leq K)$

At round t,

- player m selects $A_{m,t}$
- player m *observes* $X_{A_{m,t},t}$
- and receives the reward

$$X_{m,t} = \begin{cases} X_{A_{m,t},t} & \text{if no other player chose the same arm} \\ 0 & \text{else} \end{cases}$$

Goal:

- maximize $\sum_{m=1}^{M} \sum_{t=1}^{T} X_{m,t}$
- ... without communication between players

Cognitive radio: (OSA) sensing, attempt of transmission if no PU, possible collisions with other SUs

Multi-players bandits: algorithms

Idea: combine a good *bandit algorithm* with an *orthogonalization strategy* (collision avoidance protocol)

Example: UCB1 + ρ^{rand}. At round t each player

- has a stored rank $R_{m,t} \in \{1, \dots, M\}$
- selects the arm that has the $R_{m,t}$-largest UCB
- if a collision occurs, draws a new rank $R_{m,t+1} \sim \mathcal{U}(\{1, \dots, M\})$

Early references:

Liu and Zhao, *Distributed Learning in Multi-Armed Bandit with Multiple Players*, IEEE Trans. S. P., 2010

Anandkumar et al., *Distributed Algorithms for Learning and Cognitive Medium Access with Logarithmic Regret*, IEEE Journal on Selected Areas in Communications, 2011

Multi-players bandits: algorithms

Idea: combine a good *bandit algorithm* with an *orthogonalization strategy* (collision avoidance protocol)

Example: UCB1 + ρ^{rand}. At round t each player

- has a stored rank $R_{m,t} \in \{1, \dots, M\}$
- selects the arm that has the $R_{m,t}$-largest UCB
- if a collision occurs, draws a new rank $R_{m,t+1} \sim \mathcal{U}(\{1, \dots, M\})$

Remarks:

- M has to be known $\rightarrow$ try to estimate it
- does not handle an evolving number of devices
- is it a *fair* orthogonalization rule?
- any index policy may be used in place of UCB1

Multi-players bandits: algorithms

Idea: combine a good *bandit algorithm* with an *orthogonalization strategy* (collision avoidance protocol)

Example: UCB + ρ^{rand}. At round t each player

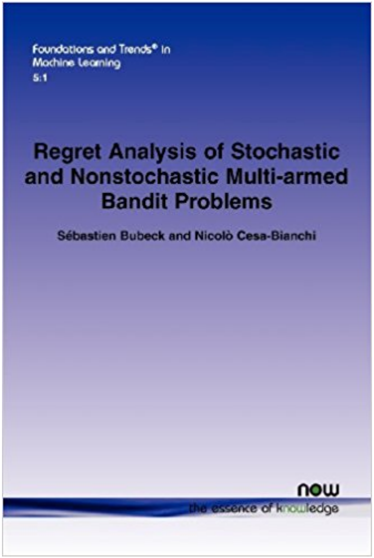
- has a stored rank $R_{m,t} \in \{1, \ldots, M\}$
- selects the arm that has the $R_{m,t}$-largest UCB
- if a collision occurs, draws a new rank $R_{t+1} \sim \mathcal{U}(\{1, \ldots, M\})$

Remarks:

- M has to be known $\rightarrow$ try to estimate it
- does not handle an evolving number of devices
- is it a *fair* orthogonalization rule?
- any index policy may be used in place of UCB1

➜ How does it perform in practice?

To read more...


Foundations and Trends® in
Machine Learning
5:1
Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems
Sébastien Bubeck and Nicolò Cesa-Bianchi
now
the essence of knowledge