A tutorial on Multi-Armed Bandit problems: Theory and Practice

Emilie Kaufmann

Imaging seminar
November 9th, 2017

Stochastic Multi-Armed Bandit model

A simple stochastic model:

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

K arms $\leftrightarrow$ K (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 sample $X_t = X_{A_t,t} \sim \nu_{A_t}$

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

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

Stochastic Multi-Armed Bandit model

A simple stochastic model:

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

K arms $\leftrightarrow$ K (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 sample $X_t = X_{A_t,t} \sim \nu_{A_t}$ (reward)

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

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

Several bandit problems

A simple stochastic model:

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

$$K \text{ arms} \leftrightarrow K \text{ (unknown) probability distribution}$$

ν_1 ν_2 ν_3 ν_4 ν_5

Several possible goals:

- find quickly the arm with largest mean (optimal exploration)

Several bandit problems

A simple stochastic model:

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

$$K \text{ arms} \leftrightarrow K \text{ (unknown) probability distribution}$$

μ_1 μ_2 μ_3 μ_4 μ_5

Several possible goals:

- find quickly the arm with largest mean (optimal exploration)

Several bandit problems

A simple stochastic model:

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

$$K \text{ arms} \leftrightarrow K \text{ (unknown) probability distribution}$$

μ_1 μ_2 μ_3 μ_4 μ_5

Several possible goals:

- find quickly the arm with largest mean (optimal exploration)
- maximize cumulated rewards $\mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$ (exploration/exploitation tradeoff)

Several bandit problems

A simple stochastic model:

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

$$K \text{ arms} \leftrightarrow K \text{ (unknown) probability distribution}$$

μ_1 μ_2 μ_3 μ_4 μ_5

Several possible goals:

- find quickly the arm with largest mean (optimal exploration)
- maximize cumulated rewards $\mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$ (exploration/exploitation tradeoff)
- (more general) learn quickly *something* about the distributions ν_k

Why Bandits ?

ν_1

ν_2

ν_3

ν_4

ν_5

Goal: maximize ones' gains in a casino ?
(HOPELESS)

Clinical trials

Historical motivation [Thompson 1933]

$\mathcal{B}(\mu_1)$ $\mathcal{B}(\mu_2)$ $\mathcal{B}(\mu_3)$ $\mathcal{B}(\mu_4)$ $\mathcal{B}(\mu_5)$

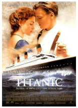
For the t-th patient in a clinical study,

- chooses a treatment A_t
- observes a response $X_t \in \{0, 1\} : \mathbb{P}(X_t = 1) = \mu_{A_t}$

Goal: identify the best treatment / maximize the number of patients healed

Online content optimization

$$ Modern motivation
(recommender systems, online advertisement, A/B Testing...)



ν_1 ν_2 ν_3 ν_4 ν_5

For the t-th visitor of a website,

- recommend a movie A_t
- observe a rating $X_t \sim \nu_{A_t}$ (e.g. $X_t \in \{1, \dots, 5\}$)

Goal: maximize the sum of ratings

Cognitive radios

Agent: a smart radio device
Arms: radio channels (frequency bands)

streams indicating channel availabilities

Channel 1	$X_{1,1}$	$X_{1,2}$	...	$X_{1,t}$	...	$X_{1,T}$
Channel 2	$X_{2,1}$	$X_{2,2}$	...	$X_{2,t}$	...	$X_{2,T}$
...	...	...	...	...	...	
Channel K	$X_{K,1}$	$X_{K,2}$	...	$X_{K,t}$	...	$X_{K,T}$

At round t, the device:

- selects channel A_t
- observes the channel availability $X_t = X_{A_t,t} = 0$ or 1

Goal: Maximize the number of sucessfull transmissions

Cognitive radios

Agent: a smart radio device
Arms: radio channels (frequency bands)

streams indicating channel availabilities

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

At round t, the device:

- selects **arm** A_t
- observes the channel availability $X_t = X_{A_t,t} = 0$ or 1

Goal: Maximize the number of sucessfull transmissions

Outline

Bandit algorithms for maximizing rewards

- First ideas
- UCB algorithms
- Bayesian algorithms

Bandit algorithms for Optimal Exploration

A Glimpse at Structured Bandit Problems

Objective

Goal: find a strategy maximizing

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

Oracle: always play the arm

$$k^* = \underset{k\in\{1,\dots,K\}}{\text{argmax}}\ \mu_k \quad \text{with mean} \quad \mu^* = \max_{k\in\{1,\dots,K\}} \mu_k.$$

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

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

Performance measure: the regret

Maximizing rewards $\leftrightarrow$ minimizing *regret*

$$R_T := T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right] = \sum_{k=1}^{K}(\mu^* - \mu_k)\mathbb{E}[N_k(T)],$$

$N_k(t)$: 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*

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

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, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$

$$R_T(\boldsymbol{\mu}) = \sum_{k=1}^{K} (\mu^* - \mu_k) \mathbb{E}_{\boldsymbol{\mu}}[N_k(T)]$$

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

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

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

  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, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$

$$R_T(\boldsymbol{\mu}) = \sum_{k=1}^{K} (\mu^* - \mu_k) \mathbb{E}_{\boldsymbol{\mu}}[N_k(T)]$$

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

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

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

  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

Some (naive) strategies

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

$\Rightarrow$ EXPLORATION

$$R(T) = \left(\frac{1}{K}\sum_{a=2}^{K}(\mu_1 - \mu_a)\right) T$$

Some (naive) strategies

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

$\Rightarrow$ EXPLORATION

$$R(T) = \left(\frac{1}{K}\sum_{a=2}^{K}(\mu_1 - \mu_a)\right) T$$

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

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

where

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

is an estimate of the unknown mean μ_k.

$\Rightarrow$ EXPLOITATION

$$\mathbb{R}(T) \geq (1-\mu_1)\times\mu_2\times(\mu_1-\mu_2)T$$

A better idea: Explore-Then-Exploit

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

- draw each arm m times
- compute the empirical best arm $\hat{k} = \text{argmax}_k \ \hat{\mu}_k(Km)$
- 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 $\hat{k} = \text{argmax}_k \; \hat{\mu}_k(Km)$
- 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, $\mu_1 > \mu_2$. $\Delta = \mu_1 - \mu_2$.

$$R_T = \Delta \times \mathbb{E}[N_2(T)]$$

$$\begin{aligned} N_2(T) &= m + (T - 2m)\mathbb{1}_{(\hat{k}=2)} \\ \mathbb{E}[N_2(T)] &\leq m + (T - 2m)\mathbb{P}\left(\hat{\mu}_1(2m) < \hat{\mu}_2(2m)\right) \\ &\leq m + T\exp\left(-\frac{m\Delta^2}{2}\right) \quad \textit{(Hoeffding's inequality)} \end{aligned}$$

A better idea: Explore-Then-Exploit

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

- draw each arm m times
- compute the empirical best arm $\hat{k} = \text{argmax}_k \, \hat{\mu}_k(Km)$
- 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, $\mu_1 > \mu_2$. $\Delta = \mu_1 - \mu_2$.

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

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

- draw each arm m times
- compute the empirical best arm $\hat{k} = \text{argmax}_k \ \hat{\mu}_k(Km)$
- 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, $\mu_1 > \mu_2$. $\Delta = \mu_1 - \mu_2$.

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

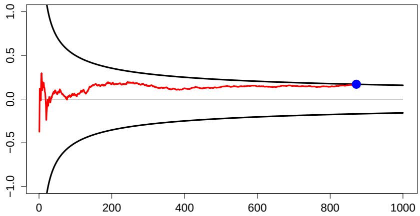
A good choice: $m = \left\lfloor \frac{2}{\Delta^2} \log\left(\frac{T\Delta^2}{2}\right) \right\rfloor$

➜ requires the knowledge of $\Delta = \mu_1 - \mu_2$!

Sequential Explore-Then-Exploit (2 arms)

- explore uniformly until the random time

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



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

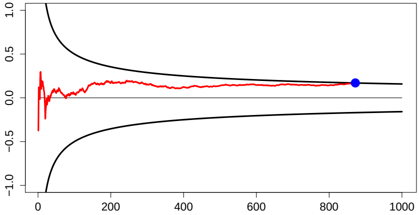
$$R_T \leq \frac{2}{\Delta}\log(T) + C\sqrt{\log(T)}.$$

➜ same regret rate, without knowing Δ

Sequential Explore-Then-Exploit (2 arms)

- explore uniformly until the random time

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



- $\hat{k} = \text{argmax}_k \ \hat{\mu}_k(\tau)$ 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

Bandit algorithms for maximizing rewards

- First ideas
- UCB algorithms
- Bayesian algorithms

Bandit algorithms for Optimal Exploration

A Glimpse at Structured Bandit Problems

The optimism principle

- For each arm k, build a confidence interval on the mean μ_k :

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

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

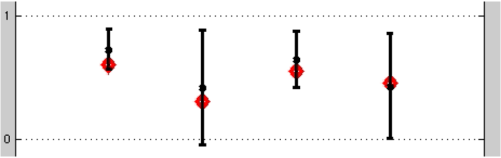


Figure: Confidence intervals on the means after t rounds

The optimism principle

- We apply the following principle:

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

(optimism in face of uncertainty)

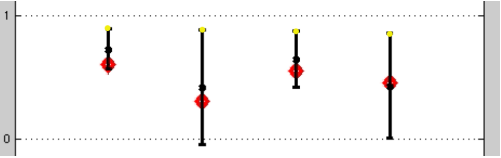


Figure: Confidence intervals on the means after t rounds

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

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

[Lai and Robbins 1985] [Agrawal 1995]

How to build the Confidence Intervals?

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

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

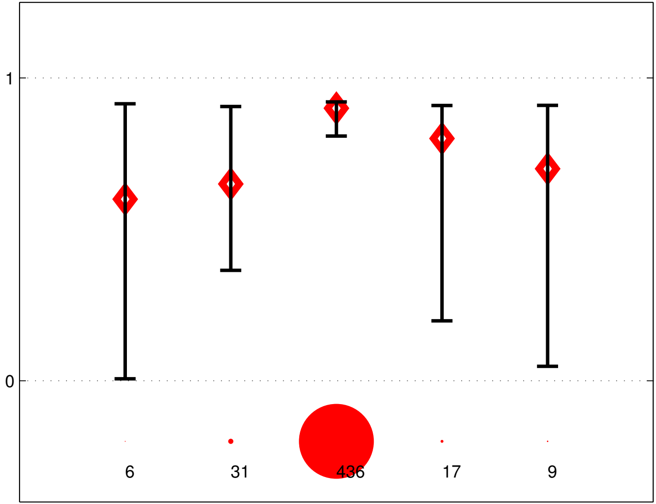
UCB1 [Auer et al. 02] chooses $A_{t+1} = \mathrm{argmax}_k \ \mathrm{U}_k(t)$ with

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

(for distributions that are 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

$$\mathrm{UCB}_k(t) = \hat{\mu}_k(t) + \sqrt{\frac{\alpha \log(t)}{N_k(t)}}$$

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

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

$$\mathbb{E}[N_k(T)] \leq \frac{\alpha}{(\mu_1 - \mu_a)^2} \log(T) + O(\sqrt{\log(T)}).$$

➜ "order-optimal" for Bernoulli distributions

$$\left[\text{Pinsker's inequality: } \; d(\mu_a, \mu_1) \geq 2(\mu_1 - \mu_a)^2\right]$$

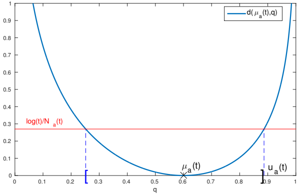
The kl-UCB algorithm

(for Bernoulli bandits, or other simple parametric families)

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

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

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



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

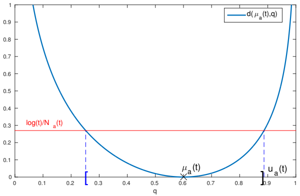
The kl-UCB algorithm

(for Bernoulli bandits, or other simple parametric families)

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

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

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



- kl-UCB is asymptotically optimal for Bernoulli bandits!

Outline

The Bayesian choice

Bernoulli bandit model $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$

- **frequentist view**: $\mu_1, \ldots, \mu_K$ are unknown parameters

➜ tools: estimators, confidence intervals

The Bayesian choice

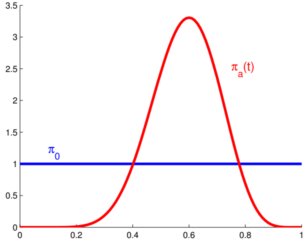
Bernoulli bandit model $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$

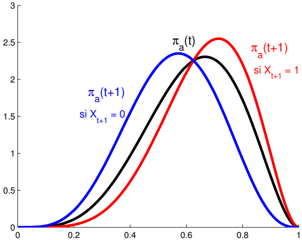
- **Bayesian view**: $\mu_1, \ldots, \mu_K$ are random variables

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

➜ tool: posterior distribution

$$\begin{aligned}\pi_k(t) &= \mathcal{L}(\mu_k | X_1, \ldots, X_t) \\ &= \mathrm{Beta}(S_k(t)+1, N_k(t)-S_k(t)+1)\end{aligned}$$

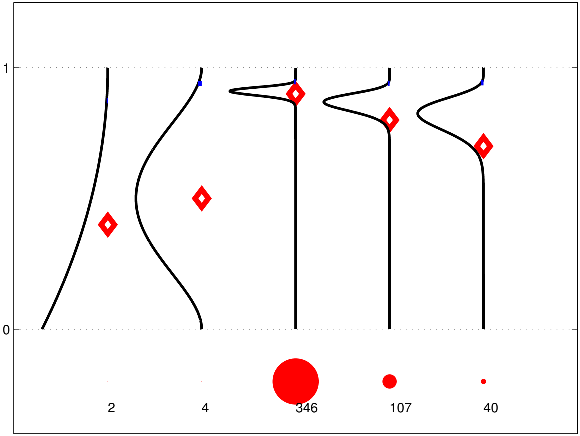




$S_k(t) = \sum_{s=1}^{t} X_s \mathbb{1}_{(A_s = k)}$ sum of the rewards from arm k

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 μ_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.

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

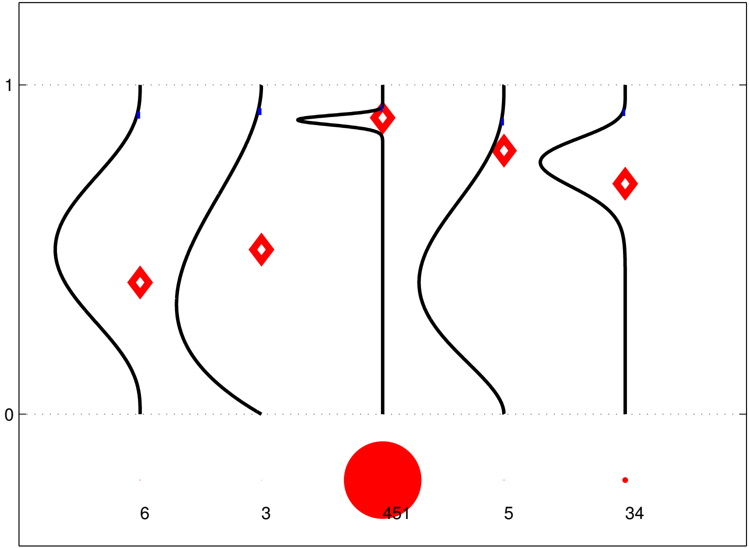
Properties:

- 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

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

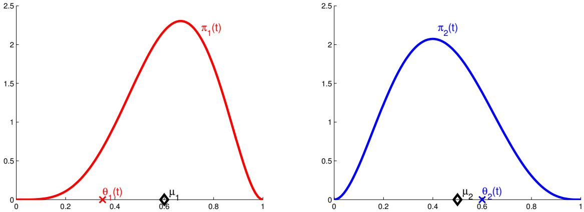


Figure: TS selects arm 2 as $\theta_2(t) \geq \theta_1(t)$

➜ 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

Outline

Bandit algorithms for maximizing rewards

- First ideas
- UCB algorithms
- Bayesian algorithms

Bandit algorithms for Optimal Exploration

A Glimpse at Structured Bandit Problems

A pure-exploration objective

Regret minimization:
maximize the number of patients healed during the trial

Alternative goal: identify as quickly as possible the best treatment
(no focus on curing patients during the study)

Two possible frameworks

The agent has to **identify the arm with highest mean** a^*
(no loss when drawing "bad" arms)

The agent

- uses a sampling strategy (A_t)
- stops at some (random) time τ
- upon stopping, recommends an arm $\hat{a}_\tau$

His goal:

Fixed-budget setting	Fixed-confidence setting
$\tau = T$ minimize $\mathbb{P}(\hat{a}_\tau \neq a^*)$	minimize $\mathbb{E}[\tau]$ $\mathbb{P}(\hat{a}_\tau \neq a^*) \leq \delta$

Fixed-budget: an elimination algorithm

SEQUENTIAL HALVING [Karnin et al. 13]

➜ $\log_2(K)$ phases of equal length, remaining arms are uniformly sampled and half of them are eliminated at the end of each phase

Initialisation: $S_0 = \{1, \dots, K\}$;

For $r = 0$ **to** $\lceil \log_2(K) \rceil - 1$, **do**

sample each arm $i \in S_r$ for $t_r = \left\lfloor \frac{T}{|S_r| \lceil \log_2(K) \rceil} \right\rfloor$ times;

let $\hat{\mu}_i^r$ be the empirical mean of arm i;

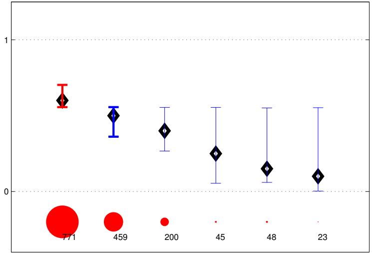
let S_{r+1} be the set of $\lceil |S_r|/2 \rceil$ arms with largest $\hat{\mu}_i^r$

Return $\hat{k}_n$ is the arm in $S_{\lceil \log_2(K) \rceil}$

Fixed-confidence: using confidence intervals

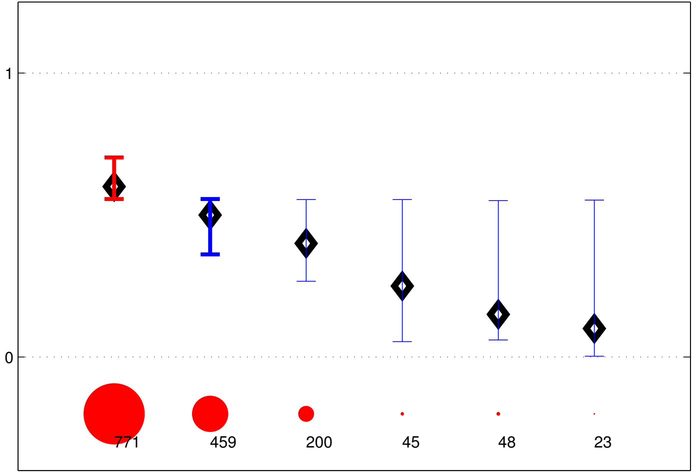
LUCB [Kalyanakrishnan et al. 12] relies on Upper And Lower confidence bounds. For KL-LUCB:

$$u_a(t) = \max\{q : N_a(t)d(\hat{\mu}_a(t), q) \leq \log(Kt/\delta)\}$$
$$\ell_a(t) = \min\{q : N_a(t)d(\hat{\mu}_a(t), q) \leq \log(Kt/\delta)\}$$



- sampling rule: $B_{t+1} = \underset{a}{\text{argmax}}\ \hat{\mu}_a(t)$, $C_{t+1} = \underset{b \neq A_{t+1}}{\text{argmax}}\ u_b(t)$
- stopping rule: $\tau = \inf\{t \in \mathbb{N} : \ell_{B_t}(t) > u_{C_t}(t)\}$

KL-LUCB in action



Theoretical garantees

$\mu_1 > \mu_2 \geq \cdots \geq \mu_K$.

- Fixed-budget setting

Theorem [Karnin et al. 13]

Sequential Halving using a budget T satisfies

$$\mathbb{P}\left(\hat{a}_T \neq 1\right) \leq 3\log_2(K)\exp\left(-\frac{T}{8H(\boldsymbol{\mu})\log_2(K)}\right)$$

with $H(\boldsymbol{\mu}) \simeq \sum_{a=2}^{K} \frac{1}{\Delta_a^2}$ and $\Delta_a = \mu_1 - \mu_a$.

- Fixed-confidence setting

Theorem [Kalyanakrishan et al.]

For well-chosen confidence intervals, LUCB is (δ)-PAC and

$$\mathbb{E}\left[\tau_\delta\right] = O\left(\left[\frac{1}{\Delta_2^2} + \sum_{a=2}^{K}\frac{1}{\Delta_a^2}\right]\log\left(\frac{1}{\delta}\right)\right)$$

The complexity of best-arm identification

Theorem [K. and Garivier, 16]

For any δ-PAC algorithm,

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geq T^*(\boldsymbol{\mu}) \log\left(\frac{1}{2.4\delta}\right),$$

where

$$T^*(\boldsymbol{\mu})^{-1} = \sup_{w \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^{K} w_a d(\mu_a, \lambda_a)\right).$$

➜ an optimal strategy satisfies $\frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(\tau)]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau]} \simeq w_a^*(\boldsymbol{\mu})$ with

$$w^*(\boldsymbol{\mu}) = \underset{w \in \Sigma_K}{\mathrm{argmax}} \inf_{\lambda \in \mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^{K} w_a d(\mu_a, \lambda_a)\right)$$

➜ tracking these optimal proportions yield a δ-PAC algorithm

such that $\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} = T^*(\boldsymbol{\mu})$.

Outline

Bandit algorithms for maximizing rewards

- First ideas
- UCB algorithms
- Bayesian algorithms

Bandit algorithms for Optimal Exploration

A Glimpse at Structured Bandit Problems

Contextual bandits

➜ incorporate *informations* about arms/agents in the model

At time t, a set of 'contexts' $\mathcal{D}_t \subset \mathbb{R}^d$ is revealed. An agent

- chooses $x_t \in \mathcal{D}_t$
- receives a reward $r_t = x_t^T\theta + \epsilon_t$.

Correlated arms: arm x_t has distribution $\mathcal{N}\left(x_t^T\theta, \sigma^2\right)$

- **Bayesian model**:

$$y_t = x_t^T\theta + \epsilon_t, \qquad \theta \sim \mathcal{N}\left(0, \kappa^2 \mathrm{I}_d\right), \qquad \epsilon_t \sim \mathcal{N}\left(0, \sigma^2\right).$$

Explicit posterior: $p(\theta|x_1, y_1, \dots, x_t, y_t) = \mathcal{N}\left(\hat{\theta}(t), \Sigma_t\right)$.

- **Thompson Sampling**

$$\begin{aligned} \tilde{\theta}(t) &\sim \mathcal{N}\left(\hat{\theta}(t), \Sigma_t\right), \\ x_{t+1} &= \underset{x \in \mathcal{D}_{t+1}}{\operatorname{argmax}}\ x^T\tilde{\theta}(t). \end{aligned}$$

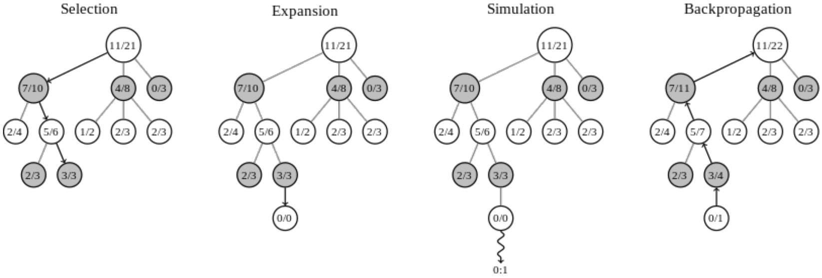
[Li et al. 12],[Agrawal & Goyal 13]

Bandits for games

To decide the next move to play:

- sequential pick trajectories in the game tree
- use (random) evaluation of some positions (playouts)

➜ How to sequentially select trajectories ?
(i.e. perform smart Monte Carlo Tree Search)



UCT algorithm [Kocsis & Szepesvari 06]: **UC**B for **T**rees
BAI-MCTS algorithms [K. & Koolen 17]

Multi-player bandits

M players simulatenously playing on the *same* MAB

At round t:

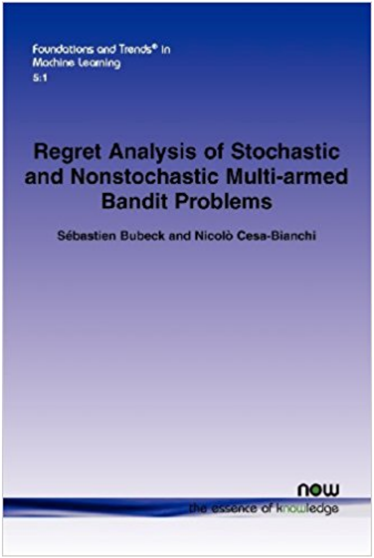
- player j selects arm $A^j(t)$
- a collision occurs for player j if another player selects the same arm
- player j receives a reward $r^j(t) = X_{A^j(t),t}\mathbb{1}_{(\overline{C^j(t)})}$

➜ simultaneously learn the quality of the channels and how to coordinate to avoid collisions and maximize global rewards

(cognitive radios: M smart devices in the same background traffic)

[Zhao et al. 10][Anandkumar et al. 11] [Besson and K. 17]

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

A new bandit game

At round t

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

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

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

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

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