Multi-armed bandit models: a tutorial

Emilie Kaufmann

CERMICS seminar,
March 30th, 2016
Multi-Armed Bandit model: general setting

\( K \) arms:

for \( a \in \{1, \ldots, K\} \), \((X_{a,t})_{t \in \mathbb{N}}\) is a stochastic process.

(unknown distributions)

**Bandit game:** a each round \( t \), an agent

- chooses an arm \( A_t \in \{1, \ldots, K\} \)
- receives a reward \( X_t = X_{A_t,t} \)

**Goal:** Build a sequential strategy

\[
A_t = F_t(A_1, X_1, \ldots, A_{t-1}, X_{t-1})
\]

maximizing

\[
\mathbb{E} \left[ \sum_{t=1}^{\infty} \alpha_t X_t \right],
\]

where \((\alpha_t)_{t \in \mathbb{N}}\) is a discount sequence. [Berry and Fristedt, 1985]
Multi-Armed Bandit model: the i.i.d. case

\( K \) independent arms:

for \( a \in \{1, \ldots, K\} \), \((X_{a,t})_{t \in \mathbb{N}}\) is i.i.d. \( \sim \nu_a \)

(unknown distributions)

**Bandit game:** a each round \( t \), an agent

- chooses an arm \( A_t \in \{1, \ldots, K\} \)
- receives a reward \( X_t \sim \nu_{A_t} \)

**Goal:** Build a sequential strategy

\[ A_t = F_t(A_1, X_1, \ldots, A_{t-1}, X_{t-1}) \]

maximizing

<table>
<thead>
<tr>
<th>Finite-horizon MAB</th>
<th>Discounted MAB</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathbb{E} \left[ \sum_{t=1}^{T} X_t \right] )</td>
<td>( \mathbb{E} \left[ \sum_{t=1}^{\infty} \alpha^{t-1} X_t \right] )</td>
</tr>
</tbody>
</table>

Emilie Kaufmann
Bandit models: a tutorial
Why MABs?

Goal: maximize ones’ gains in a casino?
(HOPELESS)
**Clinical trials:**

- Choose a treatment $A_t$ for patient $t$
- Observe a response $X_t \in \{0, 1\}$: $\mathbb{P}(X_t = 1) = \mu_{A_t}$
- **Goal:** maximize the number of patient healed

**Recommendation tasks:**

- Recommend a movie $A_t$ for visitor $t$
- Observe a rating $X_t \sim \nu_{A_t}$ (e.g. $X_t \in \{1, \ldots, 5\}$)
- **Goal:** maximize the sum of ratings
Bernoulli bandit models

\( K \) independent arms:

\[
\text{for } a \in \{1, \ldots, K\}, \quad (X_{a,t})_{t \in \mathbb{N}} \text{ is i.i.d } \sim B(\mu_a)
\]

**Bandit game:** a each round \( t \), an agent

- chooses an arm \( A_t \in \{1, \ldots, K\} \)
- receives a reward \( X_t \sim B(\mu_{A_t}) \in \{0, 1\} \)

**Goal:** maximize

<table>
<thead>
<tr>
<th>Finite-horizon MAB</th>
<th>Discounted MAB</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathbb{E} \left[ \sum_{t=1}^{T} X_t \right] )</td>
<td>( \mathbb{E} \left[ \sum_{t=1}^{\infty} \alpha^{t-1} X_t \right] )</td>
</tr>
</tbody>
</table>
Bernoulli bandit models

$K$ independent arms:

for $a \in \{1, \ldots, K\}$, $(X_{a,t})_{t \in \mathbb{N}}$ is i.i.d. $\sim \mathcal{B}(\mu_a)$

Bandit game: a each round $t$, an agent

- chooses an arm $A_t \in \{1, \ldots, K\}$
- receives a reward $X_t \sim \mathcal{B}(\mu_{A_t}) \in \{0, 1\}$

Goal: maximize

<table>
<thead>
<tr>
<th>Finite-horizon MAB</th>
<th>Discounted MAB</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbb{E} \left[ \sum_{t=1}^{T} X_t \right]$</td>
<td>$\mathbb{E} \left[ \sum_{t=1}^{\infty} \alpha^{t-1} X_t \right]$</td>
</tr>
</tbody>
</table>

Frequentist model  | Bayesian model |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_1, \ldots, \mu_K$ unknown parameters</td>
<td>$\mu_1, \ldots, \mu_K$ drawn from a prior distribution $\mu_a \sim \pi_a$</td>
</tr>
<tr>
<td>arm $a$: $(X_{a,t})_t$ i.i.d. $\sim \mathcal{B}(\mu_a)$</td>
<td>arm $a$: $(X_{a,t})_t</td>
</tr>
</tbody>
</table>
Outline

1. Bayesian bandits: a planning problem
2. Frequentist bandits: asymptotically optimal algorithms
3. Non stochastic bandits: minimax algorithms
Outline

1. Bayesian bandits: a planning problem
2. Frequentist bandits: asymptotically optimal algorithms
3. Non stochastic bandits: minimax algorithms
A Markov Decision Process

Bandit model \((\mathcal{B}(\mu_1), \ldots, \mathcal{B}(\mu_K))\)

- prior distribution: \(\mu_a \overset{\text{i.i.d}}{\sim} U([0, 1])\)
- posterior distribution: \(\pi_a^t := \mathcal{L}(\mu_a | X_1, \ldots, X_t)\)

\[
\pi_a^t = \text{Beta}\left(\underbrace{S_a(t) + 1}_{\#\text{ones}}, \underbrace{N_a(t) - S_a(t) + 1}_{\#\text{zeros}}\right)
\]

\(S_a(t)\): sum of the rewards gathered from arm \(a\) up to time \(t\)

\(N_a(t)\): number of draws of arm \(a\) up to time \(t\)

State \(\Pi^t = (\pi_a^t)_{a=1}^K\) that evolves in a MDP.
A Markov Decision Process

An example of transition:

\[
\begin{pmatrix}
    1 & 2 \\
    5 & 1 \\
    0 & 2
\end{pmatrix} \xrightarrow{A_t=2} \begin{pmatrix}
    1 & 2 \\
    6 & 1 \\
    0 & 2
\end{pmatrix} \text{ if } X_t = 1
\]

Solving a planning problem: there exists an exact solution to

- The finite-horizon MAB:
  \[
  \arg \max_{(A_t)} \mathbb{E}_{\mu \sim \pi} \left[ \sum_{t=1}^{T} X_t \right]
  \]

- The discounted MAB:
  \[
  \arg \max_{(A_t)} \mathbb{E}_{\mu \sim \pi} \left[ \sum_{t=1}^{\infty} \alpha^{t-1} X_t \right]
  \]

Optimal policy = solution to dynamic programming equations.
A Markov Decision Process

An example of transition:

\[
\begin{pmatrix}
1 & 2 \\
5 & 1 \\
0 & 2
\end{pmatrix}
\overset{A_t=2}{\rightarrow}
\begin{pmatrix}
1 & 2 \\
6 & 1 \\
0 & 2
\end{pmatrix}
\]

if \( X_t = 1 \)

Solving a planning problem: there exists an exact solution to

- The finite-horizon MAB:

\[
\arg\max_{(A_t)} \mathbb{E}_{\mu \sim \pi} \left[ \sum_{t=1}^{T} X_t \right]
\]

- The discounted MAB:

\[
\arg\max_{(A_t)} \mathbb{E}_{\mu \sim \pi} \left[ \sum_{t=1}^{\infty} \alpha^{t-1} X_t \right]
\]

Optimal policy = solution to dynamic programming equations.

**Problem:** The state space is too large!

Emilie Kaufmann
Bandit models: a tutorial
A reduction of the dimension

[Gittins 79]: the solution of the discounted MAB reduces to an index policy:

\[
A_{t+1} = \arg\max_{a=1...K} G_\alpha(\pi^t_a).
\]

- The Gittins indices:

\[
G_\alpha(p) = \sup_{\text{stopping times } \tau > 0} \frac{\mathbb{E}_{Y_t \sim B(\mu), \mu \sim p} \left[ \sum_{t=1}^\tau \alpha^{t-1} Y_t \right]}{\mathbb{E}_{Y_t \sim B(\mu), \mu \sim p} \left[ \sum_{t=1}^\tau \alpha^{t-1} \right]}
\]

"instantaneous rewards when committing to arm \( \mu \sim p \), when rewards are discounted by \( \alpha \)"
The Gittins indices

An alternative formulation:

$$G_\alpha(p) = \inf\{\lambda \in \mathbb{R} : V^*_\alpha(p, \lambda) = 0\},$$

with

$$V^*_\alpha(p, \lambda) = \sup_{\tau > 0} \mathbb{E}_{Y_t \sim \text{i.i.d.} B(\mu) \mu \sim p} \left[ \sum_{t=1}^{\tau} \alpha^{t-1}(Y_t - \lambda) \right].$$

“price worth paying for committing to arm $\mu \sim p$ when rewards are discounted by $\alpha$”
The Finite-Horizon Gittins indices: depend on the remaining time to play $r$

$$G(p, r) = \inf \{ \lambda \in \mathbb{R} : V_r^*(p, \lambda) = 0 \},$$

with

$$V_r^*(p, \lambda) = \sup_{\text{stopping times } 0 < \tau \leq r} \mathbb{E}_{\substack{Y_t \sim \text{i.i.d.} \mathcal{B}(\mu) \\ \mu \sim p}} \left[ \sum_{t=1}^{\tau} (Y_t - \lambda) \right].$$

“price worth paying for playing arm $\mu \sim p$ for at most $r$ rounds”

The Finite-Horizon Gittins algorithm

$$A_{t+1} = \arg\max_{a=1 \ldots K} G(\pi_t^a, T - t)$$

does NOT coincide with the optimal solution [Berry and Fristedt 85]... but is conjectured to be a good approximation!
Outline

1. Bayesian bandits: a planning problem
2. Frequentist bandits: asymptotically optimal algorithms
3. Non stochastic bandits: minimax algorithms
Regret minimization

\[ \mu = (\mu_1, \ldots, \mu_K) \text{ unknown parameters, } \mu^* = \max_a \mu_a. \]

- The regret of a strategy \( A = (A_t) \) is defined as

\[
R_\mu(A, T) = \mathbb{E}_\mu \left[ \mu^* T - \sum_{t=1}^{T} X_t \right]
\]

and can be rewritten

\[
R_\mu(A, T) = \sum_{a=1}^{K} (\mu^* - \mu_a) \mathbb{E}_\mu [N_a(T)].
\]

\( N_a(t) \): number of draws of arm \( a \) up to time \( t \)

Maximizing rewards ⇔ Minimizing regret

**Goal:** Design strategies that have small regret for all \( \mu \).
Optimal algorithms for regret minimization

All the arms should be drawn infinitely often!

- [Lai and Robbins, 1985]: a uniformly efficient strategy
  \((\forall \mu, \forall \alpha \in ]0, 1[, R_{\mu}(A, T) = o(T^\alpha))\) satisfies

\[
\mu_a < \mu^* \Rightarrow \lim_{T \to \infty} \inf \frac{\mathbb{E}_\mu[N_a(T)]}{\log T} \geq \frac{1}{d(\mu_a, \mu^*)},
\]

where

\[
d(\mu, \mu') = \text{KL}(\mathcal{B}(\mu), \mathcal{B}(\mu'))
\]

\[
= \mu \log \frac{\mu}{\mu'} + (1 - \mu) \log \frac{1 - \mu}{1 - \mu'}.
\]

Definition

A bandit algorithm is **asymptotically optimal** if, for every \(\mu\),

\[
\mu_a < \mu^* \Rightarrow \limsup_{T \to \infty} \frac{\mathbb{E}_\mu[N_a(T)]}{\log T} \leq \frac{1}{d(\mu_a, \mu^*)}
\]
First algorithms

- **Idea 1**: Draw each arm $T/K$ times
  
  $\Rightarrow$ EXPLORATION
First algorithms

- **Idea 1**: Draw each arm $T/K$ times
  $\Rightarrow$ **EXPLORATION**

- **Idea 2**: Always choose the empirical best arm:
  
  $$A_{t+1} = \arg \max_a \hat{\mu}_a(t)$$

  $\Rightarrow$ **EXPLOITATION**
First algorithms

- **Idea 1**: Draw each arm $T/K$ times
  $$\Rightarrow \text{EXPLORATION}$$

- **Idea 2**: Always choose the empirical best arm:
  $$A_{t+1} = \arg\max_a \hat{\mu}_a(t)$$
  $$\Rightarrow \text{EXPLOITATION}$$

- **Idea 3**: Draw the arms uniformly during $T/2$ rounds, then draw the empirical best until the end
  $$\Rightarrow \text{EXPLORATION followed EXPLOITATION}$$
First algorithms

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

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

  \[ A_{t+1} = \arg\max_a \hat{\mu}_a(t) \]

  - ⇒ **EXPLOITATION**

- **Idea 3**: Draw the arms uniformly during $T/2$ rounds, then draw the empirical best until the end
  - ⇒ **EXPLORATION followed EXPLOITATION**

**Linear regret...**
Optimistic algorithms

For each arm $a$, build a confidence interval on $\mu_a$:

$$\mu_a \leq \text{UCB}_a(t) \quad \text{w.h.p}$$

Figure: Confidence intervals on the arms at round $t$

Optimism principle:

“act as if the best possible model were the true model”

$$A_{t+1} = \arg \max_a \text{UCB}_a(t)$$
A UCB algorithm in action!
A UCB algorithm in action!
The UCB1 algorithm

UCB1 [Auer et al. 02] is based on the index

$$\text{UCB}_a(t) = \hat{\mu}_a(t) + \sqrt{\frac{\alpha \log(t)}{2N_a(t)}}$$

- Hoeffding’s inequality:

$$\mathbb{P} \left( \hat{\mu}_{a,s} + \sqrt{\frac{\alpha \log(t)}{2s}} \leq \mu_a \right) \leq \exp \left( -2s \left( \frac{\alpha \log(t)}{2s} \right) \right) = \frac{1}{t^\alpha}.$$ 

- Union bound:

$$\mathbb{P}(\text{UCB}_a(t) \leq \mu_a) \leq \mathbb{P} \left( \exists s \leq t : \hat{\mu}_{a,s} + \sqrt{\frac{\alpha \log(t)}{2s}} \leq \mu_a \right) \leq \sum_{s=1}^{t} \frac{1}{t^\alpha} = \frac{1}{t^{\alpha-1}}.$$
The UCB1 algorithm

Theorem

For every $\alpha > 2$ and every sub-optimal arm $a$, there exists a constant $C_\alpha > 0$ such that

$$\mathbb{E}_\mu [N_a(T)] \leq \frac{2\alpha}{(\mu^* - \mu_a)^2} \log(T) + C_\alpha.$$ 

It follows that

$$R_T \leq 2\alpha \left( \sum_{a \neq a^*} \frac{1}{(\mu^* - \mu_a)} \right) \log(T) + KC_\alpha.$$
Proof: 1/3

Assume \( \mu^* = \mu_1 \) and \( \mu_2 < \mu_1 \).

\[
N_2(T) = \sum_{t=0}^{T-1} \mathbb{1}(A_{t+1} = 2)
\]

\[
= \sum_{t=0}^{T-1} \mathbb{1}(A_{t+1} = 2) \cap (UCB_1(t) \leq \mu_1) + \sum_{t=0}^{T-1} \mathbb{1}(A_{t+1} = 2) \cap (UCB_1(t) > \mu_1)
\]

\[
\leq \sum_{t=0}^{T-1} \mathbb{1}(UCB_1(t) \leq \mu_1) + \sum_{t=0}^{T-1} \mathbb{1}(A_{t+1} = 2) \cap (UCB_2(t) > \mu_1)
\]
Proof: 1/3

Assume $\mu^* = \mu_1$ and $\mu_2 < \mu_1$.

$$N_2(T) = \sum_{t=0}^{T-1} 1(A_{t+1}=2)$$

$$= \sum_{t=0}^{T-1} 1(A_{t+1}=2) \cap (UCB_1(t) \leq \mu_1) + \sum_{t=0}^{T-1} 1(A_{t+1}=2) \cap (UCB_1(t) > \mu_1)$$

$$\leq \sum_{t=0}^{T-1} 1(UCB_1(t) \leq \mu_1) + \sum_{t=0}^{T-1} 1(A_{t+1}=2) \cap (UCB_2(t) > \mu_1)$$

$$\mathbb{E}[N_2(T)] \leq \sum_{t=0}^{T-1} \mathbb{P}(UCB_1(t) \leq \mu_1) + \sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = 2, UCB_2(t) > \mu_1)$$

$$\leq \underbrace{\sum_{t=0}^{T-1} \mathbb{P}(UCB_1(t) \leq \mu_1)}_{A} + \underbrace{\sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = 2, UCB_2(t) > \mu_1)}_{B}$$
Proof : 2/3

\[
E[N_2(T)] \leq \sum_{t=0}^{T-1} \mathbb{P}(UCB_1(t) \leq \mu_1) + \sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = 2, UCB_2(t) > \mu_1)
\]

- **Term A**: if \( \alpha > 2 \),

\[
\sum_{t=0}^{T-1} \mathbb{P}(UCB_1(t) \leq \mu_1) \leq 1 + \sum_{t=1}^{T-1} \frac{1}{t^{\alpha-1}} \\
\leq 1 + \zeta(\alpha - 1) := C_{\alpha}/2.
\]
Term B: 

\[
(B) = \sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = 2, \text{UCB}_2(t) > \mu_1)
\]

\[
\leq \sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = 2, \text{UCB}_2(t) > \mu_1, \text{LCB}_2(t) \leq \mu_2) + C_{\alpha}/2
\]

with

\[
\text{LCB}_2(t) = \hat{\mu}_2(t) - \sqrt{\frac{\alpha \log t}{2N_2(t)}}.
\]

(LCB_2(t) < \mu_2 < \mu_1 \leq \text{UCB}_2(t))

\[
\Rightarrow (\mu_1 - \mu_2) \leq 2 \sqrt{\frac{\alpha \log(T)}{2N_2(t)}}
\]

\[
\Rightarrow N_2(t) \leq \frac{2\alpha}{(\mu_1 - \mu_2)^2} \log(T)
\]
• **Term B:** (continued)

\[
(B) \leq \sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = 2, \text{UCB}_2(t) > \mu_1, \text{LCB}_2(t) \leq \mu_2) + C_{\alpha}/2
\]

\[
\leq \sum_{t=0}^{T-1} \mathbb{P} \left( A_{t+1} = 2, N_2(t) \leq \frac{2\alpha}{(\mu_1 - \mu_2)^2} \log(T) \right) + C_{\alpha}/2
\]

\[
\leq \frac{2\alpha}{(\mu_1 - \mu_2)^2} \log(T) + C_{\alpha}/2
\]

• **Conclusion:**

\[
\mathbb{E}[N_2(T)] \leq \frac{2\alpha}{(\mu_1 - \mu_2)^2} \log(T) + C_{\alpha}.
\]
Theorem

For every $\alpha > 2$ and every sub-optimal arm $a$, there exists a constant $C_\alpha > 0$ such that

$$E_{\mu}[N_a(T)] \leq \frac{2\alpha}{(\mu^* - \mu_a)^2} \log(T) + C_\alpha.$$
The UCB1 algorithm

Theorem

For every $\alpha > 2$ and every sub-optimal arm $a$, there exists a constant $C_\alpha > 0$ such that

$$\mathbb{E}_\mu [N_a(T)] \leq \frac{2\alpha}{(\mu^* - \mu_a)^2} \log(T) + C_\alpha.$$  

Remark:

$$\frac{2\alpha}{(\mu^* - \mu_a)^2} > 4\alpha \frac{1}{d(\mu_a, \mu^*)}$$

(UCB1 not asymptotically optimal)
The KL-UCB algorithm

- A UCB-type algorithm: \( A_{t+1} = \arg \max_a u_a(t) \)
- ... associated to the right upper confidence bounds:

\[
u_a(t) = \max \left\{ q \geq \hat{\mu}_a(t) : d(\hat{\mu}_a(t), x) \leq \frac{\log(t)}{N_a(t)} \right\},
\]

\( \hat{\mu}_a(t) \): empirical mean of rewards from arm \( a \) up to time \( t \).

[Cappé et al. 13]: KL-UCB satisfies

\[
\mathbb{E}_\mu[N_a(T)] \leq \frac{1}{d(\mu_a, \mu^*)} \log T + O(\sqrt{\log(T)}).
\]
Algorithms based on Bayesian tools can be good to solve (frequentist) regret minimization.

**Ideas:**
- Use the Finite-Horizon Gittins
- Use posterior quantiles
- Use posterior samples
Thompson Sampling

\((\pi_a^t, \ldots, \pi_K^t)\) posterior distribution on \((\mu_1, \ldots, \mu_K)\) at round \(t\).

**Algorithm: Thompson Sampling**

**Thompson Sampling** is a randomized Bayesian algorithm:

\[
\forall a \in \{1..K\}, \quad \theta_a(t) \sim \pi_a^t \\
A_{t+1} = \arg\max_a \theta_a(t)
\]

“Draw each arm according to its posterior probability of being optimal” [Thompson 1933]

Thompson Sampling is asymptotically optimal. [K., Korda, Munos 2012]
At time $t$, a set of 'contexts' $D_t \subset \mathbb{R}^d$ is revealed.

$\mathrel{=} \text{characteristics of the items to recommend}$

**The model:**
- if the context $x_t \in D_t$ is selected
- a reward $r_t = x_t^T \theta + \epsilon_t$ is received

\[
\theta \in \mathbb{R}^d \text{ = underlying preference vector}
\]

**A Bayesian model:** (with Gaussian prior)

\[
r_t = x_t^T \theta + \epsilon_t, \quad \theta \sim \mathcal{N} \left(0, \kappa^2 I_d\right), \quad \epsilon_t \sim \mathcal{N} \left(0, \sigma^2\right).
\]

Explicit posterior: $p(\theta|x_1, r_1, \ldots, x_t, r_t) = \mathcal{N} \left(\hat{\theta}(t), \Sigma_t\right)$.

**Thompson Sampling:**

\[
\tilde{\theta}(t) \sim \mathcal{N} \left(\hat{\theta}(t), \Sigma_t\right), \quad \text{and} \quad x_{t+1} = \arg\max_{x \in D_{t+1}} x^T \tilde{\theta}(t).
\]
Outline

1. Bayesian bandits: a planning problem
2. Frequentist bandits: asymptotically optimal algorithms
3. Non stochastic bandits: minimax algorithms
Minimax regret

- In stochastic bandits, we exhibited algorithm such that
  \[ \forall \mu, \ E_\mu[N_a(T)] \leq \log T/d(\mu_a, \mu^*) + o(\log T). \]

Their regret satisfy

\[
R_\mu(A, T) = \sum_{a=2}^{K} \min \left[ \frac{(\mu^* - \mu_a)}{d(\mu_a, \mu^*)} \log(T), \frac{\mu^* - \mu_a}{T} \right] + o(\log(T)).
\]

- There exist some constant \( C \) such that
  \[ \forall \mu, \ R_\mu(A, T) \leq C \sqrt{KT \log(T)}. \]

Minimax rate of the regret

\[
\inf_A \sup_\mu R_\mu(A, T) = O \left( \sqrt{KT} \right)
\]
A new bandit 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 $x_t = x_{A_t,t}$

Goal: maximize rewards, or minimize regret

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

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

at round $t$, choose

$$A_t \sim \hat{p}_t$$
Back to the bandit case: the EXP3 strategy

We don’t have access to the \((x_{a,t})\) for all \(a\)...

\[
\hat{x}_{a,t} = \frac{x_{a,t}}{\hat{p}_{a,t}} \mathbb{1}(A_t=a)
\]

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

The EXP3 strategy [Auer et al. 2003]

With \(\hat{p}_t\) the probability distribution

\[
\hat{p}_{a,t} \propto e^{\eta \left( \sum_{s=1}^{t-1} \hat{x}_{a,s} \right)}
\]

at round \(t\), choose

\[
A_t \sim \hat{p}_t
\]
The EXP3 strategy [Auer et al. 2003]

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

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

at round $t$, choose

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

[Bubeck and Cesa-Bianchi 12] EXP3 with

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

satisfies

$$R(\text{EXP3}, T) \leq \sqrt{2 \log K \sqrt{KT}}$$

Remarks:

- almost the same guarantees for $\eta_t = \sqrt{\frac{\log(K)}{Kt}}$
- extra exploration is needed to have high probability results
Conclusions

Under different assumptions, different types of strategies to achieve an exploration-exploitation tradeoff in bandit models:

**Index policies:**
- Gittins indices
- UCB-type algorithms

**Randomized algorithms:**
- Thompson Sampling
- Exponential weights

More complex bandit models not covered today: restless bandits, contextual bandits, combinatorial bandits...
Beyond regret minimization

Under different assumptions, different types of strategies to achieve an exploration-exploitation tradeoff in bandit models:

**Index policies:**
- Gittins indices
- UCB-type algorithms

**Randomized algorithms:**
- Thompson Sampling
- Exponential weights

More complex bandit models not covered today:
restless bandits, contextual bandits, combinatorial bandits...
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)
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)

Additionnaly to the sampling strategy \( (A_t) \), one needs

- a stopping rule \( \tau \) (stopping time)
- a recommendation rule \( \hat{a}_\tau \)

such that, for some risk parameter \( \delta \in ]0, 1[ \),

\[
P(\hat{a}_\tau \neq a^*) \leq \delta \quad \text{and} \quad E[\tau] \text{ is as small as possible.}
\]
An algorithm: KL-LUCB [K., Kalyanakrishnan 13]

An algorithm based on Upper and Lower confidence bounds

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

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

- sampling rule: \( A_{t+1} = \arg\max_a \hat{\mu}_a(t), \quad B_{t+1} = \arg\max_a u_b(t) \) \( b \neq A_{t+1} \)
- stopping rule: \( \tau = \inf \{ t \in \mathbb{N} : \ell_{A_t}(t) > u_{B_t}(t) \} \)
- recommendation rule: \( \hat{a}_\tau = \arg\max_{a=1, \ldots, K} \hat{\mu}_a(\tau) \)
KL-LUCB for finding the $m$ best arms

Emilie Kaufmann

Bandit models: a tutorial
The complexity of best-arm identification

Theorem [K. and Garivier, 16]

For any $\delta$-PAC algorithm,

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

where

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

⇒ an optimal strategy satisfies

$$\frac{\mathbb{E}_\mu[N_a(\tau)]}{\mathbb{E}_\mu[\tau]} \simeq w^*_a(\mu)$$

with

$$w^*(\mu) = \arg \max_{w \in \Sigma_K} \inf_{\lambda \in \text{Alt}(\mu)} \left(\sum_{a=1}^{K} w_a d(\mu_a, \lambda_a)\right).$$

⇒ tracking these optimal proportions yield a $\delta$-PAC algorithm such that

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

**Bayesian bandits:**

**Stochastic and non-stochastic bandits:**

**Best arm identification:**