Bandit models: a tutorial

Emilie Kaufmann

Gdt COS,
December 3rd, 2015

Multi-Armed Bandit model: general setting

K arms:

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

(**unknown** distributions)

Bandit game: a each round t, an agent

- chooses an arm $A_t \in \{1, \dots, K\}$
- receives a reward $X_t = X_{A_t,t}$

Goal: Build a sequential strategy

$$A_t = F_t(A_1, X_1, \dots, A_{t-1}, X_{t-1})$$

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:

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

(**unknown** distributions)

Bandit game: a each round t, an agent

- chooses an arm $A_t \in \{1, \dots, K\}$
- receives a reward $X_t \sim \nu_{A_t}$

Goal: Build a sequential strategy

$$A_t = F_t(A_1, X_1, \dots, A_{t-1}, X_{t-1})$$

maximizing

Finite-horizon MAB	Discounted MAB
$\mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$	$\mathbb{E}\left[\sum_{t=1}^{\infty} \alpha^{t-1} X_t\right]$

Why MABs?

ν_1 ν_2 ν_3 ν_4 ν_5

Goal: maximize ones' gains in a casino ?

(HOPELESS)

Why MABs? Real motivations

Clinical trials:

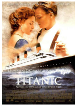




$\mathcal{B}(\mu_1)$ $\mathcal{B}(\mu_2)$ $\mathcal{B}(\mu_3)$ $\mathcal{B}(\mu_4)$ $\mathcal{B}(\mu_5)$

- 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:



ν_1 ν_2 ν_3 ν_4 ν_5

- 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:

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

Bandit game: a each round t, an agent

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

Goal: maximize

Finite-horizon MAB	Discounted MAB
$\mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$	$\mathbb{E}\left[\sum_{t=1}^{\infty} \alpha^{t-1} X_t\right]$

Bernoulli bandit models

K independent arms:

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

Bandit game: a each round t, an agent

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

Goal: maximize

Finite-horizon MAB	Discounted MAB
$\mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$	$\mathbb{E}\left[\sum_{t=1}^{\infty} \alpha^{t-1} X_t\right]$

Frequentist model	**Bayesian model**
$\mu_1, \dots, \mu_K$ unknown parameters	$\mu_1, \dots, \mu_K$ drawn from a prior distribution : $\mu_a \sim \pi_a$
arm a: $(X_{a,t})_t \overset{\text{i.i.d.}}{\sim} \mathcal{B}(\mu_a)$	arm a: $(X_{a,t})_t \lvert \boldsymbol{\mu} \overset{\text{i.i.d.}}{\sim} \mathcal{B}(\mu_a)$

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

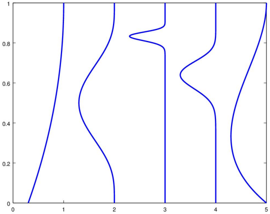
A Markov Decision Process

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

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

$$\pi_a^t = \mathrm{Beta}\Big(\underbrace{S_a(t)}_{\#ones} + 1, \underbrace{N_a(t) - S_a(t)}_{\#zeros} + 1\Big)$$

$S_a(t)$: sum of the rewards gathered from 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.

An example of transition:

$$\begin{pmatrix} 1 & 2 \\ 5 & 1 \\ 0 & 2 \end{pmatrix} \xrightarrow{A_t = 2} \begin{pmatrix} 1 & 2 \\ 6 & 1 \\ 0 & 2 \end{pmatrix} \textit{ if } X_t = 1$$

Solving a planning problem: there exists an exact solution to

- The finite-horizon MAB:

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

- The discounted MAB:

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

Optimal policy = solution to dynamic programming equations.

A Markov Decision Process

An example of transition:

$$\begin{pmatrix} 1 & 2 \\ 5 & 1 \\ 0 & 2 \end{pmatrix} \xrightarrow{A_t=2} \begin{pmatrix} 1 & 2 \\ 6 & 1 \\ 0 & 2 \end{pmatrix} \textit{ if } X_t = 1$$

Solving a planning problem: there exists an exact solution to

- The finite-horizon MAB:

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

- The discounted MAB:

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

Optimal policy = solution to dynamic programming equations.

Problem: The state space is too large !

A reduction of the dimension

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

$$A_{t+1} = \underset{a=1\dots K}{\text{argmax}} \; G_\alpha(\pi_a^t).$$

- **The Gittins indices**:

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

"instantaneous rewards when commiting to arm $\mu \sim p$, when rewards are discounted by α"

An alternative formulation:

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

with

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

"price worth paying for commiting to arm $\mu \sim p$ when rewards are discounted by α"

Gittins indices for finite horizon

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

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

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

The Finite-Horizon Gittins algorithm

$$A_{t+1} = \underset{a=1\dots K}{\text{argmax}}\ G(\pi_a^t, T-t)$$

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

$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ unknown parameters, $\mu^* = \max_a \mu_a$.

- The **regret** of a strategy $\mathcal{A} = (A_t)$ is defined as

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

and can be rewritten

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

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

Maximizing rewards $\Leftrightarrow$ Minimizing regret

Goal: Design strategies that have small regret for all $\boldsymbol{\mu}$.

Optimal algorithms for regret minimization

All the arms should be drawn infinitely often !

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

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

where

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

Definition

A bandit algorithm is **asymptotically optimal** if, for every $\boldsymbol{\mu}$,

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

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:

$$A_{t+1} = \underset{a}{\text{argmax}}\ \hat{\mu}_a(t)$$

$\Rightarrow$ EXPLOITATION

First algorithms

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

$\Rightarrow$ EXPLORATION

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

$$A_{t+1} = \underset{a}{\text{argmax}}\ \hat{\mu}_a(t)$$

$\Rightarrow$ EXPLOITATION

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

$\Rightarrow$ EXPLORATION followed EXPLOITATION

First algorithms

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

$\Rightarrow$ EXPLORATION

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

$$A_{t+1} = \underset{a}{\text{argmax}}\ \hat{\mu}_a(t)$$

$\Rightarrow$ EXPLOITATION

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

$\Rightarrow$ EXPLORATION followed EXPLOITATION

Linear regret...

Optimistic algorithms

- For each arm a, build a confidence interval on μ_a :

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

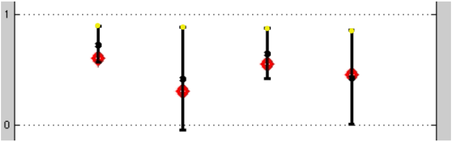


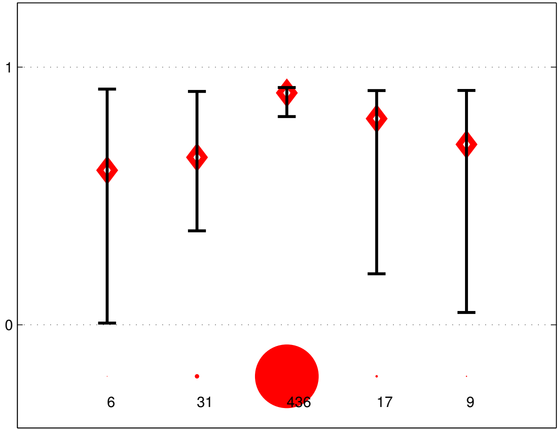
Figure : Confidence intervals on the arms at round t

- Optimism principle:

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

$$A_{t+1} = \underset{a}{\arg\max}\ \mathrm{UCB}_a(t)$$

A UCB algorithm in action !



The UCB1 algorithm

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

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

- 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:

$$\begin{aligned}\mathbb{P}(\mathrm{UCB}_a(t) \leq \mu_a) &\leq \mathbb{P}\left(\exists s \leq t : \hat{\mu}_{a,s} + \sqrt{\frac{\alpha \log(t)}{2s}} \leq \mu_a\right) \\ &\leq \sum_{s=1}^{t} \frac{1}{t^\alpha} = \frac{1}{t^{\alpha-1}}.\end{aligned}$$

The UCB1 algorithm

Theorem

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

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

The UCB1 algorithm

Theorem

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

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

Remark:

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

(UCB1 not asymptotically optimal)

Proof : 1/3

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

$$
\begin{aligned}
N_2(T) &= \sum_{t=0}^{T-1} \mathbb{1}_{(A_{t+1}=2)} \\
&= \sum_{t=0}^{T-1} \mathbb{1}_{(A_{t+1}=2)\cap(\mathrm{UCB}_1(t)\leq\mu_1)} + \sum_{t=0}^{T-1} \mathbb{1}_{(A_{t+1}=2)\cap(\mathrm{UCB}_1(t)>\mu_1)} \\
&\leq \sum_{t=0}^{T-1} \mathbb{1}_{(\mathrm{UCB}_1(t)\leq\mu_1)} + \sum_{t=0}^{T-1} \mathbb{1}_{(A_{t+1}=2)\cap(\mathrm{UCB}_2(t)>\mu_1)}
\end{aligned}
$$

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

$$\begin{aligned}
N_2(T) &= \sum_{t=0}^{T-1} \mathbb{1}_{(A_{t+1}=2)} \\
&= \sum_{t=0}^{T-1} \mathbb{1}_{(A_{t+1}=2)\cap(\mathrm{UCB}_1(t)\leq\mu_1)} + \sum_{t=0}^{T-1} \mathbb{1}_{(A_{t+1}=2)\cap(\mathrm{UCB}_1(t)>\mu_1)} \\
&\leq \sum_{t=0}^{T-1} \mathbb{1}_{(\mathrm{UCB}_1(t)\leq\mu_1)} + \sum_{t=0}^{T-1} \mathbb{1}_{(A_{t+1}=2)\cap(\mathrm{UCB}_2(t)>\mu_1)}
\end{aligned}$$

$$\mathbb{E}[N_2(T)] \leq \underbrace{\sum_{t=0}^{T-1} \mathbb{P}(\mathrm{UCB}_1(t) \leq \mu_1)}_{A} + \underbrace{\sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = 2, \mathrm{UCB}_2(t) > \mu_1)}_{B}$$

Proof : 2/3

$$\mathbb{E}[N_2(T)] \leq \underbrace{\sum_{t=0}^{T-1} \mathbb{P}(\mathrm{UCB}_1(t) \leq \mu_1)}_{A} + \underbrace{\sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = 2, \mathrm{UCB}_2(t) > \mu_1)}_{B}$$

- **Term** A: if $\alpha > 2$,

$$\begin{aligned}\sum_{t=0}^{T-1} \mathbb{P}(\mathrm{UCB}_1(t) \leq \mu_1) &\leq 1 + \sum_{t=1}^{T-1} \frac{1}{t^{\alpha-1}} \\ &\leq 1 + \zeta(\alpha - 1) := C_\alpha/2.\end{aligned}$$

- **Term B:**

$$
\begin{aligned}
(B) &= \sum_{t=0}^{T-1} \mathbb{P}\left(A_{t+1} = 2, \mathrm{UCB}_2(t) > \mu_1\right) \\
&\leq \sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = 2, \mathrm{UCB}_2(t) > \mu_1, \mathrm{LCB}_2(t) \leq \mu_2) + C_\alpha/2
\end{aligned}
$$

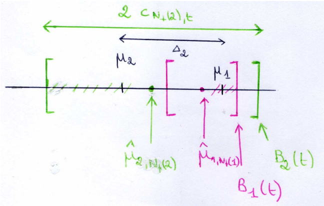
with

$$\mathrm{LCB}_2(t) = \hat{\mu}_2(t) - \sqrt{\frac{\alpha \log t}{2N_2(t)}}.$$

$$(\mathrm{LCB}_2(t) < \mu_2 < \mu_1 \leq \mathrm{UCB}_2(t))$$

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

$$\Rightarrow \quad N_2(t) \leq \frac{2\alpha}{(\mu_1 - \mu_2)^2} \log(T)$$



Proof : 3/3

- **Term B:** (continued)

$$\begin{aligned} (B) &\leq \sum_{t=0}^{T-1} \mathbb{P}(A_{t+1} = 2, \mathrm{UCB}_2(t) > \mu_1, \mathrm{LCB}_2(t) \leq \mu_2) + C_\alpha/2 \\ &\leq \sum_{t=0}^{T-1} \mathbb{P}\left(A_{t+1} = 2, N_2(t) \leq \frac{2\alpha}{(\mu_1 - \mu_2)^2}\log(T)\right) + C_\alpha/2 \\ &\leq \frac{2\alpha}{(\mu_1 - \mu_2)^2}\log(T) + C_\alpha/2 \end{aligned}$$

- **Conclusion**:

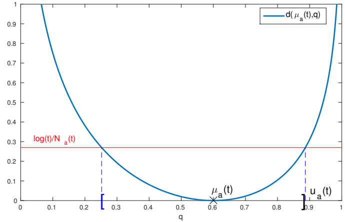
$$\mathbb{E}[N_2(T)] \leq \frac{2\alpha}{(\mu_1 - \mu_2)^2}\log(T) + C_\alpha.$$

The KL-UCB algorithm

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

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

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

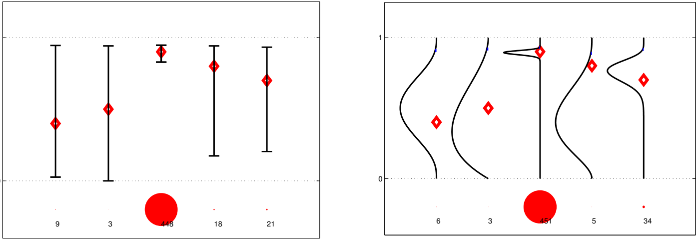


[Cappé et al. 13]: KL-UCB satisfies

$$\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] \leq \frac{1}{\mathrm{d}(\mu_a, \mu^*)} \log T + O(\sqrt{\log(T)}).$$

Bayesian algorithms for regret minimization?

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

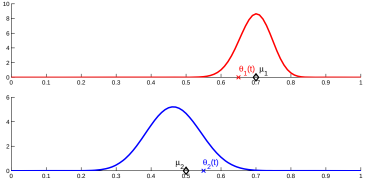
$(\pi_a^t, .., \pi_K^t)$ posterior distribution on $(\mu_1, .., \mu_K)$ at round t.

Algorithm: Thompson Sampling

Thompson Sampling is a randomized Bayesian algorithm:

$$\forall a \in \{1..K\}, \quad \theta_a(t) \sim \pi_a^t$$
$$A_{t+1} = \text{argmax}_a \ \theta_a(t)$$

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



Thompson Sampling is asymptotically optimal.

[K.,Korda,Munos 2012]

Outline

Minimax regret

In stochastic (Bernoulli) bandits, we exhibited algorithm satisfying

$$\forall \boldsymbol{\mu}, \ R_{\boldsymbol{\mu}}(\mathcal{A}, T) = \underbrace{\left(\sum_{a=1}^{K} \frac{(\mu^* - \mu_a)}{\mathrm{d}(\mu_a, \mu^*)} \right)}_{\substack{\text{problem-dependent term,} \\ \text{can be large}}} \log(T) + o(\log(T)).$$

For those algorithms, one can also prove that, for some constant C,

$$\forall \boldsymbol{\mu}, \ R_{\boldsymbol{\mu}}(\mathcal{A}, T) \leq \underbrace{C\sqrt{KT\log(T)}}_{\text{problem-independent bound}}.$$

Minimax rate of the regret

$$\inf_{\mathcal{A}} \sup_{\boldsymbol{\mu}} R_{\boldsymbol{\mu}}(\mathcal{A}, T) = O\left(\sqrt{KT}\right)$$

Removing the stochastic assumption...

A new bandit game: at round t

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

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

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

Goal: maximize rewards, or minimize regret

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

Full information: Exponential Weighted Forecaster

The full-information game: at round t

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

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

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

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

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

Theoretical results

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

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

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...

Books and surveys

Bayesian bandits:

- *Bandit Problems. Sequential allocation of experiments*.
  Berry and Fristedt. Chapman and Hall, 1985.
- *Multi-armed bandit allocation indices*.
  Gittins, Glazebrook, Weber. Wiley, 2011.

Stochastic and non-stochastic bandits:

- *Regret analysis of Stochastic and Non-stochastic Bandit Problems*. Bubeck and Cesa-Bianchi.
  Fondations and Trends in Machine Learning, 2012.
- *Prediction, Learning and Games.*
  Cesa-Bianchi and Lugosi. Cambridge University Press, 2006.