Multi-Armed Bandits : a Bayesian view

Emilie Kaufmann

EURO PhD school, July 2022

Historical perspective

1952 Robbins, formulation of the MAB problem

1985 Lai and Robbins : lower bound, first asymptotically optimal algorithm

1987 Lai, asymptotic regret of kl-UCB

1995 Agrawal, UCB algorithms

1995 Katehakis and Robbins, a UCB algorithm for Gaussian bandits

2002 Auer et al : UCB1 with finite-time regret bound

2009 UCB-V, MOSS...

2011,13 Cappé et al : finite-time regret bound for kl-UCB

Historical perspective

1933 Thompson : a Bayesian mechanism for clinical trials

1952 Robbins, formulation of the MAB problem

1956 Bradt et al, Bellman : optimal solution of a Bayesian MAB problem

1979 Gittins : first Bayesian index policy

1985 Lai and Robbins : lower bound, first asymptocally optimal algorithm

1985 Berry and Fristedt : Bandit Problems, a survey on the Bayesian MAB

1987 Lai, asymptotic regret of kl-UCB + study of its Bayesian regret

1995 Agrawal, UCB algorithms

1995 Katehakis and Robbins, a UCB algorithm for Gaussian bandits

2002 Auer et al : UCB1 with finite-time regret bound

2009 UCB-V, MOSS...

2010 Thompson Sampling is re-discovered

2011,13 Cappé et al : finite-time regret bound for kl-UCB

2012,13 Thompson Sampling is asymptotically optimal

Recap : the multi-armed bandit setup

$\nu = (\nu_1, \dots, \nu_K)$ set of arms
ν_a has mean μ_a

At round t, an agent :

- chooses an arm A_t (based on past observation)
- receives a reward $R_t \sim \nu_{A_t}$

$(Y_{a,s})_{s \in \mathbb{N}^\star}$: stream of successive rewards from arm a, i.i.d. under ν_a
$R_t = Y_{a,N_a(t)}$ where $N_a(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = a)$

Goal : Maximize $\mathbb{E}\left[\sum_{t=1}^{T} R_t\right] \leftrightarrow$ minimize the regret

$$\mathcal{R}_\nu(T) = \mathbb{E}_\nu\left[\sum_{t=1}^{T}(\mu_\star - \mu_{A_t})\right] = \sum_{a=1}^{K}(\mu_\star - \mu_a)\mathbb{E}_\nu[N_a(T)]$$

Recap : the multi-armed bandit setup

$\nu = (\nu^{\mu_1}, \dots, \nu^{\mu_K})$ set of arms (parametric distributions)
ν^{μ_a} has mean μ_a

At round t, an agent :

- chooses an arm A_t (based on past observation)
- receives a reward $R_t \sim \nu_{A_t}$

$(Y_{a,s})_{s\in\mathbb{N}^\star}$: stream of successive rewards from arm a, i.i.d. under ν_a
$R_t = Y_{a,N_a(t)}$ where $N_a(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = a)$

Goal : Maximize $\mathbb{E}\left[\sum_{t=1}^{T} R_t\right] \leftrightarrow$ minimize the regret

$$\mathcal{R}_{\boldsymbol{\mu}}(T) = \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{T}(\mu_\star - \mu_{A_t})\right] = \sum_{a=1}^{K}(\mu_\star - \mu_a)\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]$$

Two probabilistic models

$\nu_{\boldsymbol{\mu}} = (\nu^{\mu_1}, \dots, \nu^{\mu_K}) \in (\mathcal{P})^K$.

- Two probabilistic models

Frequentist model	**Bayesian model**
$\mu_1, \dots, \mu_K$ unknown parameters	$\mu_1, \dots, \mu_K$ drawn from a prior distribution : $\boldsymbol{\mu} \sim \pi$
arm a : $(Y_{a,s})_s \overset{\text{i.i.d.}}{\sim} \nu^{\mu_a}$	arm a : $(Y_{a,s})_s \lvert \boldsymbol{\mu} \overset{\text{i.i.d.}}{\sim} \nu^{\mu_a}$

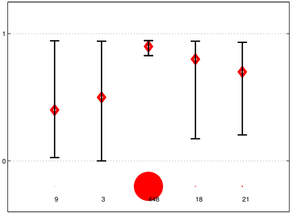
Frequentist regret (regret)	Bayesian regret (Bayes risk)
$\mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}, T) = \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{T} (\mu_\star - \mu_{A_t})\right]$	$\mathrm{R}^\pi(\mathcal{A}, T) = \mathbb{E}_{\boldsymbol{\mu}\sim\pi}\left[\sum_{t=1}^{T} (\mu_\star - \mu_{A_t})\right] = \int \mathcal{R}_{\boldsymbol{\mu}}(\mathcal{A}, T) d\pi(\boldsymbol{\mu})$

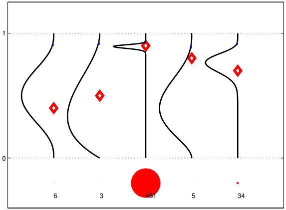
Particular case : product prior $\pi = (\pi_1 \otimes \dots \otimes \pi_K)$

Two types of algorithms

- Two types of tools to build bandit algorithms :

Frequentist tools	Bayesian tools
MLE estimators of the means Confidence Intervals	Posterior distributions $\pi_a^t = \mathcal{L}(\mu_a \| Y_{a,1}, \ldots, Y_{a,N_a(t)})$

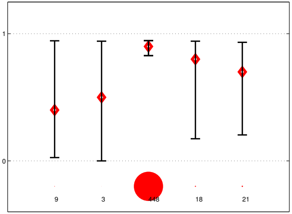


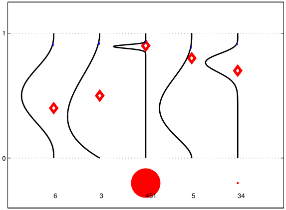


Two types of algorithms

- Two types of tools to build bandit algorithms :

Frequentist tools	Bayesian tools
MLE estimators of the means Confidence Intervals	Posterior distributions $\pi_a^t = \mathcal{L}(\mu_a \| Y_{a,1}, \dots, Y_{a,N_a(t)})$





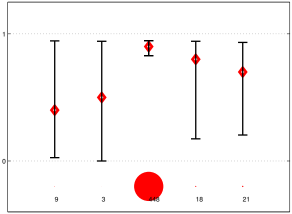
Remark : Tools $\neq$ objective !

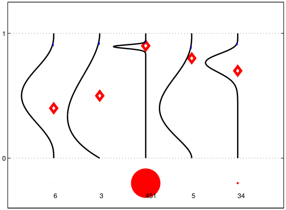
➜ we can analyze the (frequentist) regret of Bayesian algorithms

Two types of algorithms

- Two types of tools to build bandit algorithms :

Frequentist tools	Bayesian tools
MLE estimators of the means Confidence Intervals	Posterior distributions $\pi_a^t = \mathcal{L}(\mu_a \| Y_{a,1}, \ldots, Y_{a,N_a(t)})$





Remark : Tools $\neq$ objective !

➜ we can analyze the Bayes risk of frequentist algorithms

Example : Bernoulli bandits

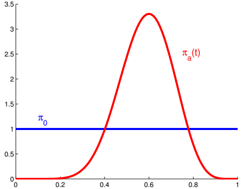
Bernoulli bandit model $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$

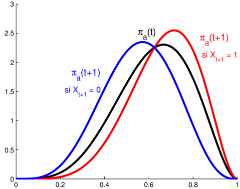
- **Bayesian view** : $\mu_1, \dots, \mu_K$ are random variables
  prior distribution : $\mu_a \overset{\text{i.i.d.}}{\sim} \mathcal{U}([0,1])$

➜ posterior distribution :

$$\begin{aligned}\pi_a(t) &= \mathcal{L}\left(\mu_a | R_1, \dots, R_t\right) \\ &= \mathrm{Beta}\Big(\underbrace{S_a(t)}_{\#ones}+1, \underbrace{N_a(t) - S_a(t)}_{\#zeros}+1\Big)\end{aligned}$$

$N_a(t) = \sum_{s=1}^{t} \mathbb{1}_{(A_s=a)}$ number of observations from arm a
$S_a(t) = \sum_{s=1}^{t} R_s \mathbb{1}_{(A_s=a)}$ sum of the rewards from arm a





Example : Gaussian bandits

Gaussian bandit model $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, known variance σ^2

- **Bayesian view** : $\mu_1, \ldots, \mu_K$ are random variables
  prior distribution : $\mu_a \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \kappa^2)$

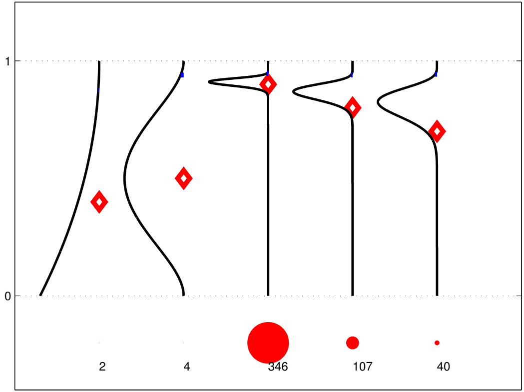
➜ posterior distribution :

$$\begin{aligned} \pi_a(t) &= \mathcal{L}\left(\mu_a | R_1, \ldots, R_t\right) \\ &= \mathcal{N}\left(\frac{S_a(t)}{N_a(t) + \frac{\sigma^2}{\kappa^2}}, \frac{\sigma^2}{N_a(t) + \frac{\sigma^2}{\kappa^2}}\right) \end{aligned}$$

$N_a(t) = \sum_{s=1}^{t} \mathbb{1}_{(A_s = a)}$ number of observations from arm a
$S_a(t) = \sum_{s=1}^{t} R_s \mathbb{1}_{(A_s = a)}$ sum of the rewards from arm a

Bayesian algorithms

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



Outline

1. Bayesian Optimal Solution and Gittins Indices
2. A Bayesian view on optimism
3. Thompson Sampling
4. Re-sampling methods

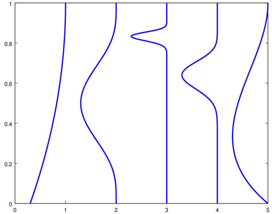
Bayesian optimal solution

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

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

The posterior distribution is fully summarized by a matrix containing the two parameters of the Beta distribution for each arm.

$$\Pi^t = \begin{pmatrix} 1 & 3 \\ 4 & 4 \\ 14 & 5 \\ 6 & 3 \\ 2 & 4 \end{pmatrix}$$



"State" Π^t

A Markov Decision Process

After each arm selection A_t, we receive a reward R_t such that

$$\mathbb{P}\left(R_t = 1 | \Pi^{t-1} = \Pi, A_t = a\right) = \underbrace{\frac{\Pi^{t-1}(a,1)}{\Pi^{t-1}(a,1) + \Pi^{t-1}(a,2)}}_{\text{mean of } \pi_a(t-1)}$$

and the posterior gets updated :

$$\begin{aligned} \Pi^t(A_t,1) &= \Pi^{t-1}(A_t,1) + R_t \\ \Pi^t(A_t,2) &= \Pi^{t-1}(A_t,2) + (1 - R_t) \end{aligned}$$

Example of transition :

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

➜ Markov Decision Process with $\mathcal{S} = \{\text{possible posteriors } \Pi\}$, $\mathcal{A} = \{1, \dots, K\}$ and **known** dynamics

A Markov Decision Process

After each arm selection A_t, we receive a reward R_t such that

$$\mathbb{P}\left(R_t = 1 | \Pi^{t-1} = \Pi, A_t = a\right) = \underbrace{\frac{\Pi^{t-1}(a,1)}{\Pi^{t-1}(a,1) + \Pi^{t-1}(a,2)}}_{\text{mean of } \pi_a(t-1)}$$

and the posterior gets updated :

$$\begin{aligned} \Pi^t(A_t,1) &= \Pi^{t-1}(A_t,1) + R_t \\ \Pi^t(A_t,2) &= \Pi^{t-1}(A_t,2) + (1 - R_t) \end{aligned}$$

Example of transition :

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

➜ Markov Decision Process with $\mathcal{S} = \{\text{possible posteriors } \Pi\}$, $\mathcal{A} = \{1, \dots, K\}$ and **known** dynamics

Solving the MDP

Solving the Bayesian bandit problem (i.e. minimizing Bayes risk)
$\leftrightarrow$ maximizing rewards in some Markov Decision Process

There exists an exact solution to

- The finite-horizon MAB :

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

- The discounted MAB :

$$\underset{(A_t)}{\text{argmax}} \ \mathbb{E}_{\boldsymbol{\mu}\sim\pi}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R_t\right]$$

[Berry and Fristedt, *Bandit Problems*, 1985]

Optimal solution : solution to dynamic programming equations.

Problem : The state space is very large (if not infinite)

$\rightsquigarrow$ often intractable

Optimal solution : tractability

For the Finite-Horizon case, the optimal policy can be computed using backwards induction : $V_{T+1}^\star = 0$ and

$$V_h^\star(\Pi) = \max_{a\in\{1,\dots,K\}} \left(\mathbb{E}_{\boldsymbol{\mu}\sim\Pi}\left[\mu_a\right] + \mathbb{E}_{\substack{X\sim\nu^{\mu_a}\\ \mu_a\sim\boldsymbol{\mu}}}\left[V_{h+1}^\star\left(\Pi_{a,X}\right)\right]\right)$$

$\Pi_{a,X}$: new posterior obtained from Π after an additional reward X from arm a

Bernoulli bandits :

$$\begin{aligned} V_h^\star(\Pi) = \quad & \max_{a\in\{1,\dots,K\}} \left(\frac{\Pi(a,1)}{\Pi(a,1)+\Pi(a,2)} + \frac{\Pi(a,1)}{\Pi(a,1)+\Pi(a,2)} V_{h+1}^\star\left(\Pi_{a,1}\right)\right. \\ & \left. + \frac{\Pi(a,2)}{\Pi(a,1)+\Pi(a,2)} V_{h+1}^\star\left(\Pi_{a,0}\right)\right) \end{aligned}$$

➜ requires a lot of memory !

Gittins indices

[Gittins, 1979] : for product priors, the solution of the discounted MAB

$$\underset{(A_t)}{\text{argmax}} \ \mathbb{E}_{\boldsymbol{\mu}\sim\pi}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R_t\right]$$

is an **index policy** :

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

- **The Gittins indices** :

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

with

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

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

Gittins indices for finite horizon ?

The solution of the finite horizon MAB

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

is NOT an index policy. [Berry and Fristedt, 1985]

- **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

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

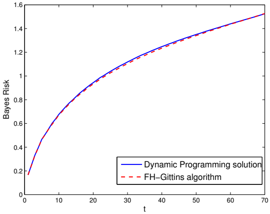
"price worth paying for playing arm $\mu \sim p$ for at most r rounds"

Finite Horizon Gittins algorithm

FH-Gittins algorithm :

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

does NOT coincide with the Bayesian optimal solution but is conjectured to be a good approximation !



- good performance in terms of (frequentist) regret as well
- logarithmic regret proved for Gaussian bandits [Lattimore, 2016]
- Gittins indices remain costly compared to UCB [Nino-Mora, 2011]

Outline

Approximations of the FH-Gittins indices

- [Burnetas and Katehakis, 2003] : when r is large,

$$G(\pi_a(t-1), r) \simeq \max\left\{q : N_a(t) \times \mathrm{kl}\left(\hat{\mu}_a(t), q\right) \leq \log\left(\frac{r}{N_a(t)}\right)\right\}$$

- [Lai, 87] : the index policy associated to

$$I_a(t) = \max\left\{q : N_a(t) \times \mathrm{kl}\left(\hat{\mu}_a(t), q\right) \leq \log\left(\frac{T}{N_a(t)}\right)\right\}$$

is a good approximation of the Bayesian solution for large T.

➜ looks like the kl-UCB index, with a different exploration rate...

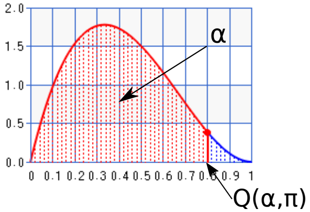
Bayes-UCB

- $\Pi_0 = (\pi_1(0), \dots, \pi_K(0))$ be a prior distribution over $(\mu_1, ..., \mu_K)$
- $\Pi_t = (\pi_1(t), \dots, \pi_K(t))$ be the posterior distribution over the means $(\mu_1, ..., \mu_K)$ after t observations

Bayes-UCB selects at time $t+1$

$$A_{t+1} = \underset{a=1,\dots,K}{\operatorname{argmax}} \; Q\left(1 - \frac{1}{t(\log t)^c}, \pi_a(t)\right)$$

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



Bayes-UCB

- $\Pi_0 = (\pi_1(0), \dots, \pi_K(0))$ be a prior distribution over $(\mu_1, ..., \mu_K)$
- $\Pi_t = (\pi_1(t), \dots, \pi_K(t))$ be the posterior distribution over the means $(\mu_1, ..., \mu_K)$ after t observations

Bayes-UCB selects at time $t+1$

$$A_{t+1} = \underset{a=1,\dots,K}{\operatorname{argmax}} \; Q\left(1 - \frac{1}{t(\log t)^c}, \pi_a(t)\right)$$

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

Bernoulli reward with uniform prior :

- $\pi_a(0) \overset{i.i.d}{\sim} \mathcal{U}([0,1]) = \text{Beta}(1,1)$
- $\pi_a(t) = \text{Beta}(S_a(t) + 1, N_a(t) - S_a(t) + 1)$

Bayes-UCB

- $\Pi_0 = (\pi_1(0), \dots, \pi_K(0))$ be a prior distribution over $(\mu_1, ..., \mu_K)$
- $\Pi_t = (\pi_1(t), \dots, \pi_K(t))$ be the posterior distribution over the means $(\mu_1, ..., \mu_K)$ after t observations

Bayes-UCB selects at time $t+1$

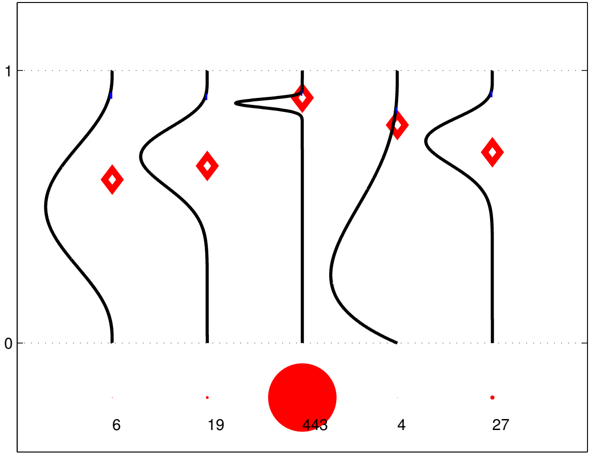
$$A_{t+1} = \underset{a=1,...,K}{\operatorname{argmax}} \; Q\left(1 - \frac{1}{t(\log t)^c}, \pi_a(t)\right)$$

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

Gaussian rewards with Gaussian prior :

- $\pi_a(0) \overset{i.i.d}{\sim} \mathcal{N}(0, \kappa^2)$
- $\pi_a(t) = \mathcal{N}\left(\frac{S_a(t)}{N_a(t)+\sigma^2/\kappa^2}, \frac{\sigma^2}{N_a(t)+\sigma^2/\kappa^2}\right)$

Bayes UCB in action



Theoretical guarantees

- Bayes-UCB is asymptotically optimal for Bernoulli rewards

Theorem [Kaufmann et al., 2012a]

Let $\epsilon > 0$. The Bayes-UCB algorithm using a uniform prior over the arms and parameter $c \geq 5$ satisfies

$$\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] \leq \frac{1+\epsilon}{\mathrm{kl}(\mu_a, \mu_\star)} \log(T) + o_{\epsilon,c}\left(\log(T)\right).$$

Why ? posterior quantile $\simeq$ kl-UCB index : $\tilde{u}_a(t) \leq q_a(t) \leq u_a(t)$ where

$$u_a(t) = \max\left\{q : \mathrm{kl}\left(\frac{S_a(t)}{N_a(t)}, q\right) \leq \frac{\log(t) + c\log(\log(t))}{N_a(t)}\right\}$$

$$\tilde{u}_a(t) = \max\left\{q : \mathrm{kl}\left(\frac{S_a(t)}{N_a(t)+1}, q\right) \leq \frac{\log\left(\frac{t}{N_a(t)+2}\right) + c\log(\log(t))}{(N_a(t)+1)}\right\}$$

Outline

Historical perspective

1933 Thompson : in the context of clinical trial with two treatments, the allocation of a treatment should be some increasing function of its posterior probability to be optimal

2010 Thompson Sampling rediscovered under different names

Bayesian Learning Automaton [Granmo, 2010]

Randomized probability matching [Scott, 2010]

2011 An empirical evaluation of Thompson Sampling : an efficient algorithm, beyond simple bandit models

[Chapelle and Li, 2011]

2012 First (logarithmic) regret bound for Thompson Sampling

[Agrawal and Goyal, 2012]

2012 Thompson Sampling is asymptotically optimal for Bernoulli bandits

[Kaufmann et al., 2012b, Agrawal and Goyal, 2013]

2013- Many successful uses of Thompson Sampling beyond Bernoulli bandits (contextual bandits, reinforcement learning)

Thompson Sampling

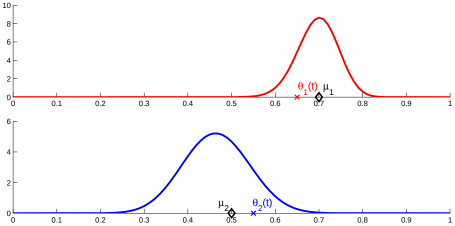
Two equivalent interpretations :

- "select an arm at random according to its probability of being the best"
- "draw a possible bandit model from the posterior distribution and act optimally in this sampled model"

$\neq$ optimistic

Thompson Sampling : a randomized Bayesian algorithm

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



Thompson Sampling is asymptotically optimal

Problem-dependent regret

$$\forall \epsilon > 0, \quad \mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] \leq (1+\epsilon)\frac{1}{\mathrm{kl}(\mu_a, \mu_\star)} \log(T) + o_{\mu,\epsilon}(\log(T)).$$

This results holds :

- for Bernoulli bandits, with a uniform prior [Kaufmann et al., 2012b, Agrawal and Goyal, 2013]
- for Gaussian bandits, with Gaussian prior [Agrawal and Goyal, 2017]
- for exponential family bandits, with Jeffrey's prior [Korda et al., 2013]

Problem-independent regret [Agrawal and Goyal, 2017]

For Bernoulli and Gaussian bandits, Thompson Sampling satisfies

$$\mathcal{R}_{\boldsymbol{\mu}}(\texttt{TS}, T) = O\left(\sqrt{KT\log(T)}\right).$$

Understanding Thompson Sampling

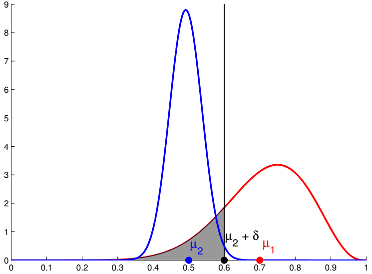
- a key ingredient in the analysis of [Kaufmann et al., 2012b]

Proposition

There exists constants $b = b(\mu) \in (0,1)$ and $C_b < \infty$ such that

$$\sum_{t=1}^{\infty} \mathbb{P}\left(N_1(t) \leq t^b\right) \leq C_b.$$

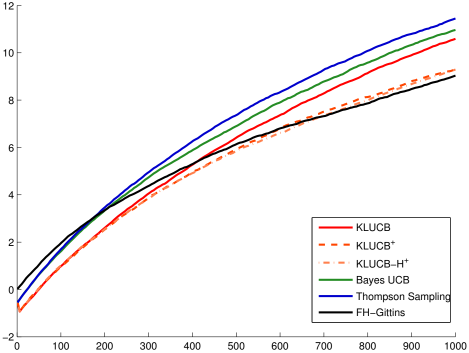
$$\left\{N_1(t) \leq t^b\right\} = \{\text{there exists a time range of length at least } t^{1-b} - 1 \text{ with no draw of arm 1 }\}$$



Practical performance

- Short horizon, $T = 1000$

2 arms Bernoulli bandit problem
$\mu_1 = 0.2, \mu_2 = 0.25$

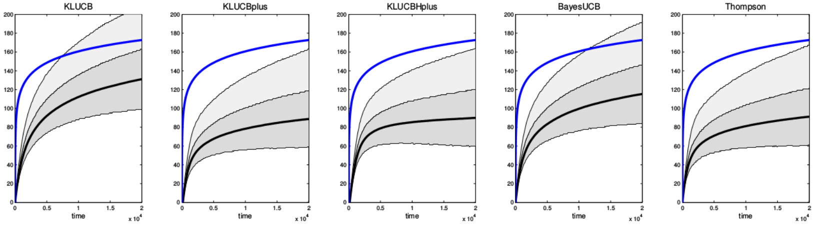


Regret as a function of time
(averaged over $N = 10000$ runs)

Practical performance

- Long horizon, $T = 20000$

10 arms Bernoulli bandit problem
$\mu = [0.1\ 0.05\ 0.05\ 0.05\ 0.02\ 0.02\ 0.02\ 0.01\ 0.01\ 0.01]$



Regret as a function of time
(average over $N = 50000$ runs)

Outline

Non parametric algorithms

Thompson Sampling relies on a parametric assumption to maintain a posterior distribution

- Gaussian rewards with known variance : TS with Gaussian prior
- Bernoulli rewards* : TS with Beta prior

Idea : replace the posterior sampling step by a non-parametric history-resampling method

*A binarization trick can be used to handle more general bounded rewards

Perturbed History Exploration

First idea : Non-parameteric Bootstrap

- $\mathcal{H}_{a,t} = (Y_{a,1}, \dots, Y_{a,N_a(t)})$: history of collected rewards from arm a
- sample $N_a(t)$ rewards from $\mathcal{H}_{a,t}$ with replacement, and average them to define an index $B_a(t)$
- $A_{t+1} = \text{argmax}_a \, B_a(t)$

[Kveton et al., 2019b] : linear regret even for two Bernoulli arms

➜ possible fix : Perturbing the history

Perturbed History Exploration (PHE)

$B_a(t)$ is the empirical means of the rewards in $\mathcal{H}_{a,t}$ and $a \times N_a(t)$ fake rewards drawn iid from $\mathcal{B}(1/2)$

➜ $a > 2$: logarithmic regret for bounded rewards in $[0, 1]$ [Kveton et al., 2019a]

Non Parametric Thompson Sampling

Context : rewards bounded in $[0, B]$
Idea : random re-weighting of the augmented history

[Riou and Honda, 2020]

Index of arm a after t rounds

- $\mathcal{H}_{a,t} = (Y_{a,1}, \dots, Y_{a,N_a(t)}, B)$: history of collected rewards from arm a augmented by the upper bound B on the support
- $w_{a,t} \sim \mathrm{Dir}(\underbrace{1, \dots, 1}_{N_a(t)+1})$ a random probability vector

$$B_a(t) = \sum_{s=1}^{N_a(t)} w_{a,t}(s) Y_{a,s} + B w_{a,N_a(t)+1}$$

Non Parametric Thompson Sampling

Context : rewards bounded in $[0, B]$
Idea : random re-weighting of the augmented history

[Riou and Honda, 2020]

Index of arm a after t rounds

- $\mathcal{H}_{a,t} = (Y_{a,1}, \dots, Y_{a,N_a(t)}, B)$: history of collected rewards from arm a augmented by the upper bound B on the support
- $w_{a,t} \sim \mathrm{Dir}(\underbrace{1, \dots, 1}_{N_a(t)+1})$ a random probability vector

$$B_a(t) = \text{mean}\left(\tilde{F}_{a,t}\right) \text{ where } \tilde{F}_{a,s} = \sum_{s=1}^{N_a(t)} w_{a,t}(s)\delta_{Y_{a,s}} + w_{a,N_a(t)+1}\delta_B$$

(perturbed CDF view)

Non Parameteric Thompson Sampling

Let $\mathcal{B}$ be the set of distributions that are supported on $[0, B]$.

Theorem [Riou and Honda, 2020]

On an instance $\nu = (\nu_1, \dots, \nu_K)$ such that $\nu_a \in \mathcal{B}$ for all a.

$$\mathcal{R}_\nu(\text{NPTS}, T) \leq \sum_{a:\mu_a<\mu_\star} \frac{\Delta_a \log T}{\mathcal{K}_{\text{inf}}(\nu_a, \mu_\star)} + o(\log T)\ .$$

where $\mathcal{K}_{\text{inf}}(\nu, \mu) = \inf \left\{\mathrm{KL}\left(\nu, \nu'\right) : \nu' \in \mathcal{B} : \mathbb{E}_{X\sim\nu'}[X] \geq \mu\right\}$.

➜ matching the lower bound of [Burnetas and Katehakis, 1996] for general (possibly non-parametric) reward distributions

A sub-sampling alternative

Idea : perform fair comparisons between pairs of arms (duels)
[Baransi et al., 2014, Chan, 2020, Baudry et al., 2020]

Sub-Sampling Duelling Algorithms (SDA) use a *round-based* structure

1. Find the *leader* : arm with largest number of observations
2. Organize $K - 1$ *duels* : *leader* vs *challengers*.
3. Draw a set of arms : *winning challengers* xor *leader*.

How do duels work ?

- challenger : compute $\hat{\mu}_c$, the empirical mean
- leader : compute $\tilde{\mu}_\ell$, the mean of a *sub-sample* of the same size as the history of the challenger.
- challenger wins if $\hat{\mu}_c \geq \tilde{\mu}_\ell$

Random Block SDA

Input of SDA : how to sub-sample n elements from N ?

- Random-Block Sampling (RB-SDA) : return a block of size n starting from random $n_0 \sim \mathcal{U}([1, N-n])$

7.6	-4	0.7	1.4	3.1	0.1	-1.2

Theorem [Baudry et al., 2020]

RB-SDA is asymptotically optimal for any bandit model whose rewards belong to an exponential family (e.g. Bernoulli, Gaussian with known variance, Poisson, Exponential).

Random Block SDA

Input of SDA : how to sub-sample n elements from N ?

- Random-Block Sampling (RB-SDA) : return a block of size n starting from random $n_0 \sim \mathcal{U}([1, N-n])$

7.6	-4	0.7	1.4	3.1	0.1	-1.2

Theorem [Baudry et al., 2020]

RB-SDA is asymptotically optimal for any bandit model whose rewards belong to an exponential family (e.g. Bernoulli, Gaussian with known variance, Poisson, Exponential).

... but it can fail for some other distributions

Practical performance

Average Regret on $N = 10000$ random instances with $K = 10$ arms

- **Bernoulli arms**

T	TS (Beta)	PHE	SSMC	RB-SDA
100	**13.8**	16.7	16.5	**14.8**
1000	**27.8**	39.5	34.2	**31.8**
10000	**45.8**	72.3	55.0	**51.1**
20000	**52.2**	85.6	61.9	**57.7**

- **Gaussian arms**

T	TS (Gaussian)	SSMC	RB-SDA
100	41.2	40.6	**38.1**
1000	76.4	76.2	**70.4**
10000	118.5	120.1	**111.8**
20000	132.6	135.1	**125.7**

Conclusion

Bayesian (inspired) algorithms

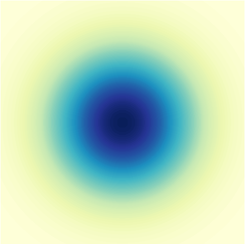
- are competitive alternative to optimistic approaches (but their analysis is generally harder)
- are flexible algorithms that can be used whenever a prior/posterior pair is available
- ... and even beyond

Bayesian algorithms

- can be extended to more complex (e.g. contextual) bandit models
- and to exploration strategies for reinforcement learning

[Osband et al., 2013, Tiapkin et al., 2022]

Thompson Sampling for RL



MDP $\boldsymbol{M}$ is drawn from some prior distribution ν_0.

$\nu_t \in \Delta(\mathcal{M})$: posterior distribution over the set of MDPs

Optimism	**Posterior Sampling**
Set of possible MDPs	Posterior distribution over MDPs
Compute the optimistic MDP	Sample from the posterior distribution

Posterior Sampling for RL

Algorithm 1: PSRL in episodic MDPs

Input : Prior distribution ν_0

for $t = 1, 2, \dots$ **do**

  $s_1 \sim \rho$ \\\\ get the starting state of episode t

  Sample $\widetilde{M}_t \sim \nu_{t-1}$ \\\\ sample an MDP from the current posterior distribution

  Compute $\tilde{\pi}^t$ an optimal policy for $\widetilde{M}_t$

  for $h = 1, \dots, H$ **do**

    $a_h = \tilde{\pi}^t_h(s_h)$ \\\\ choose next action according to $\tilde{\pi}^t$

    $r_h, s_{h+1} = \text{step}(s_h, a_h)$

  end

  Compute ν_t based on ν_{t-1} and $\{(s_h, a_h, r_h, s_{h+1})\}_{h=1}^{H}$

end

[Strens, 2000, Osband et al., 2013, Agrawal and Jia, 2017]

Agrawal, S. and Goyal, N. (2012).
Analysis of Thompson Sampling for the multi-armed bandit problem.
In *Proceedings of the 25th Conference On Learning Theory*.

Agrawal, S. and Goyal, N. (2013).
Further Optimal Regret Bounds for Thompson Sampling.
In *Proceedings of the 16th Conference on Artificial Intelligence and Statistics*.

Agrawal, S. and Goyal, N. (2017).
Near-optimal regret bounds for thompson sampling.
J. ACM, 64(5) :30 :1–30 :24.

Agrawal, S. and Jia, R. (2017).
Optimistic posterior sampling for reinforcement learning : worst-case regret bounds.
In *Advances in Neural Information Processing Systems (NIPS)*.

Baransi, A., Maillard, O., and Mannor, S. (2014).
Sub-sampling for multi-armed bandits.
In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML / PKDD*.

Baudry, D., Kaufmann, E., and Maillard, O.-A. (2020).
Sub-sampling for Efficient Non-Parametric Bandit Exploration.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Berry, D. and Fristedt, B. (1985).
Bandit Problems. Sequential allocation of experiments.

Chapman and Hall.

Burnetas, A. and Katehakis, M. (1996).
Optimal adaptive policies for sequential allocation problems.
Advances in Applied Mathematics, 17(2) :122–142.

Burnetas, A. and Katehakis, M. (2003).
Asymptotic Bayes Analysis for the finite horizon one armed bandit problem.
Probability in the Engineering and Informational Sciences, 17 :53–82.

Chan, H. P. (2020).
The multi-armed bandit problem : An efficient nonparametric solution.
The Annals of Statistics, 48(1).

Chapelle, O. and Li, L. (2011).
An empirical evaluation of Thompson Sampling.
In *Advances in Neural Information Processing Systems*.

Gittins, J. (1979).
Bandit processes and dynamic allocation indices.
Journal of the Royal Statistical Society, Series B, 41(2) :148–177.

Granmo, O. (2010).
Solving two-armed Bernoulli Bandit Problems using a Bayesian Learning Automaton.
International Journal of Intelligent Computing and Cybernetics, 3(2) :207–234.

Kaufmann, E., Cappé, O., and Garivier, A. (2012a).

On Bayesian Upper-Confidence Bounds for Bandit Problems.
In *Proceedings of the 15th conference on Artificial Intelligence and Statistics*.

Kaufmann, E., Korda, N., and Munos, R. (2012b).
Thompson Sampling : an Asymptotically Optimal Finite-Time Analysis.
In *Proceedings of the 23rd conference on Algorithmic Learning Theory*.

Korda, N., Kaufmann, E., and Munos, R. (2013).
Thompson Sampling for 1-dimensional Exponential family bandits.
In *Advances in Neural Information Processing Systems*.

Kveton, B., Szepesvári, C., Ghavamzadeh, M., and Boutilier, C. (2019a).
Perturbed-history exploration in stochastic multi-armed bandits.
In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*.

Kveton, B., Szepesvári, C., Vaswani, S., Wen, Z., Lattimore, T., and Ghavamzadeh, M. (2019b).
Garbage in, reward out : Bootstrapping exploration in multi-armed bandits.
In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Lattimore, T. (2016).
Regret analysis of the finite-horizon gittins index strategy for multi-armed bandits.
In *Conference On Learning Theory (COLT)*.

Nino-Mora, J. (2011).
Computing a Classic Index for Finite-Horizon Bandits.

INFORMS Journal of Computing, 23(2) :254–267.

Osband, I., Van Roy, B., and Russo, D. (2013).
(More) Efficient Reinforcement Learning Via Posterior Sampling.
In *Advances in Neural Information Processing Systems*.

Riou, C. and Honda, J. (2020).
Bandit algorithms based on thompson sampling for bounded reward distributions.
In *Algorithmic Learning Theory (ALT)*.

Scott, S. (2010).
A modern Bayesian look at the multi-armed bandit.
Applied Stochastic Models in Business and Industry, 26 :639–658.

Strens, M. (2000).
A Bayesian Framework for Reinforcement Learning.
In *ICML*.

Tiapkin, D., Belomestny, D., Moulines, E., Naumov, A., Samsonov, S., Tang, Y., Valko, M., and Ménard, P. (2022).
From dirichlet to rubin : Optimistic exploration in RL without bonuses.
In *International Confernece on Machine Learning (ICML)*.