On Bayesian algorithms for sequential resource allocation

Emilie Kaufmann,

joint work with O. Cappé, A. Garivier, N. Korda and R. Munos.

SIGMA seminar,
November 16th, 2015

The multi-armed bandit model (MAB)

K arms $=$ K probability distributions (ν_a has mean μ_a)

ν_1 ν_2 ν_3 ν_4 ν_5

At round t, an agent

- chooses arm A_t
- observes reward $X_t \sim \nu_{A_t}$

$\mathcal{A} = (A_t)$ is his strategy or bandit algorithm :

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

Goal: maximize the rewards obtained during T interactions

$\Leftrightarrow$ minimize regret:

$$\mathbb{E}\left[T(\max_a \mu_a) - \sum_{t=1}^{T} X_t\right] = \mathbb{E}\left[\sum_{t=1}^{T}(\mu^* - \mu_{A_t})\right]$$

Modern motivation: recommendation tasks

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

Initial motivation: clinical trials





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

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: maximize the number of patient healed during the study

Our setup: exponential family bandit model


ν_{θ_1}


ν_{θ_2}


ν_{θ_3}


ν_{θ_4}

ν_{θ_5}

$\nu_{\theta_1}, \dots, \nu_{\theta_K}$ belong to a one-dimensional exponential family:

$$\mathcal{P} = \{\nu_\theta, \theta \in \Theta : \nu_\theta \text{ has a density } f_\theta(x) = \exp(\theta x - b(\theta))\}$$

- ν_θ can be parametrized by its mean $\mu = \dot{b}(\theta)$: $\nu^\mu := \nu_{\dot{b}^{-1}(\mu)}$

For a given exponential family $\mathcal{P}$,

$$d_{\mathcal{P}}(\mu, \mu') := \text{KL}(\nu^\mu, \nu^{\mu'}) = \mathbb{E}_{X \sim \nu^\mu}\left[\log \frac{d\nu^\mu}{d\nu^{\mu'}}(X)\right]$$

is the KL-divergence between the distributions of mean μ and μ'.

Bernoulli case: $(\theta = \log \frac{\mu}{1-\mu}, \;\; b(\theta) = \log(1 + e^\theta)\ )$

$$d(\mu, \mu') = \text{KL}(\mathcal{B}(\mu), \mathcal{B}(\mu')) = \mu \log \frac{\mu}{\mu'} + (1-\mu) \log \frac{1-\mu}{1-\mu'}.$$

Outline

Outline

A frequentist or a Bayesian model?

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

- Two probabilistic modelings

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: $(Y_{a,s})_s \overset{\text{i.i.d.}}{\sim} \nu^{\mu_a}$	arm a: $(Y_{a,s})_s \lvert \boldsymbol{\mu} \overset{\text{i.i.d.}}{\sim} \nu^{\mu_a}$

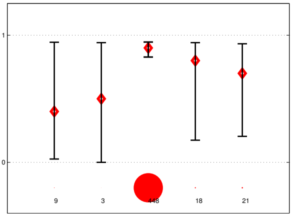
- The regret can be computed in each case

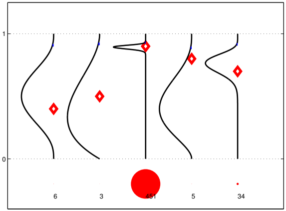
Frequentist regret (regret)	Bayesian regret (Bayes risk)
$R_T(\mathcal{A}, \boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=1}^T (\mu^* - \mu_{A_t})\right]$	$\mathcal{R}_T(\mathcal{A}, \pi) = \mathbb{E}_{\boldsymbol{\mu} \sim \pi}\left[\sum_{t=1}^T (\mu^* - \mu_{A_t})\right] = \int R_T(\mathcal{A},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

Frequentist and Bayesian algorithms

- Two types of tools to build bandit algorithms:

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





- Today:

Algorithms based on Bayesian tools
for solving (frequentist) regret minimization

Optimal algorithms for regret minimization

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

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

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

- [Lai and Robbins, 1985]:

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

Definition

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

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

Towards optimal index policies

- An index policy is of the form

$$A_{t+1} = \underset{a=1\dots K}{\arg\max}\, I_a(t)$$

$I_a(t)$: index that depends on the past observations from arm a,

$$Y_{a,1}, \dots, Y_{a,N_a(t)}.$$

Towards optimal index policies

- An index policy is of the form

$$A_{t+1} = \underset{a=1\dots K}{\arg\max}\, I_a(t)$$

$I_a(t)$: index that depends on the past observations from arm a,

$$Y_{a,1}, \dots, Y_{a,N_a(t)}.$$

- A first (bad) idea: $A_{t+1} = \arg\max_a \ \hat{\mu}_a(t)$

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

Towards optimal index policies

- An index policy is of the form

$$A_{t+1} = \underset{a=1\dots K}{\arg\max}\, I_a(t)$$

$I_a(t)$: index that depends on the past observations from arm a,

$$Y_{a,1}, \dots, Y_{a,N_a(t)}.$$

- A first (bad) idea: $A_{t+1} = \arg\max_a \ \hat{\mu}_a(t)$

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

- A better idea: $A_{t+1} = \arg\max_a \ \text{UCB}_a(t)$

$\text{UCB}_a(t)$: an upper-confidence bound on μ_a

Example: ([Auer et al. 02], Bernoulli case)

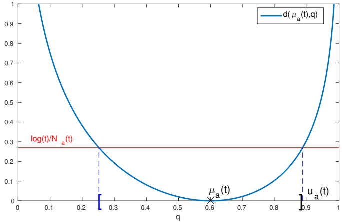
$$UCB_a(t) = \hat{\mu}_a(t) + \sqrt{\frac{2\log(t)}{N_a(t)}}.$$

The KL-UCB algorithm

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

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

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



[Cappé et al. 13]: KL-UCB satisfies, for $c \geq 5$,

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

WANTED!

Algorithms that are asymptotically optimal but also

- more efficient in practice
- easier to implement
- easier to generalize beyond exponential family bandits

Our answer:

Go Bayesian !

Outline

There exists an exact solution to Bayes risk minimization:

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

Why? The history of the game can be summarized by a posterior matrix, that evolves in a Markov Decision Process.
$\Rightarrow$ optimal policy = solution to dynamic programming equations.

Example: Bernoulli bandit model $\nu_{\boldsymbol{\mu}} = (\mathcal{B}(\mu_1), \dots, \mathcal{B}(\mu_K))$

- $\mu_a \sim \mathcal{U}([0,1])$
- $\pi_a^t = \text{Beta}(\#|\text{ones observed}| + 1, \#|\text{zeros observed}| + 1)$

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

There exists an exact solution to Bayes risk minimization:

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

Why? The history of the game can be summarized by a posterior matrix, that evolves in a Markov Decision Process.
$\Rightarrow$ optimal policy = solution to dynamic programming equations.

Example: Bernoulli bandit model $\nu_{\boldsymbol{\mu}} = (\mathcal{B}(\mu_1), \dots, \mathcal{B}(\mu_K))$

- $\mu_a \sim \mathcal{U}([0,1])$
- $\pi_a^t = \text{Beta}(\#|\text{ones observed}| + 1, \#|\text{zeros observed}| + 1)$

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

INTRACTABLE !

Gittins' solution

[Gittins 79]: the solution of the discounted MAB,

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

is an index policy:

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

[Gittins 79]: the solution of the discounted MAB,

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

is an index policy:

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

In the undiscounted case: the Finite-Horizon Gittins algorithm

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

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

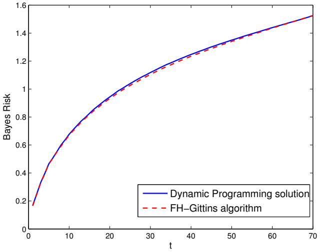
$$V_\lambda^*(p,r) = \sup_{0\leq\tau\leq r} \mathbb{E}_{\substack{Y_t \overset{\text{i.i.d}}{\sim} \nu^\mu \\ \mu\sim\pi}}\left[\sum_{t=1}^{\tau}(Y_t-\lambda)\right]$$

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

The FH-Gittins algorithm

FH-Gittins...

- does NOT coincide with the optimal solution of the undiscounted MAB ([Berry, Fristedt 1985]) but it is conjectured to be a good approximation

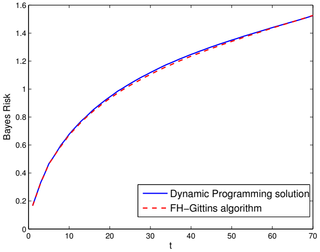


- displays good performance in terms of regret as well !

The FH-Gittins algorithm

FH-Gittins...

- does NOT coincide with the optimal solution of the undiscounted MAB ([Berry, Fristedt 1985]) but it is conjectured to be a good approximation



- displays good performance in terms of regret as well !

INDICES ARE HARD TO COMPUTE...

Approximating the FH-Gittins indices

- [Burnetas and Katehakis, 03]: when n is large,

$$G(\pi_a^t, n) \simeq \max\left\{q \geq \hat{\mu}_a(t), N_a(t) d\left(\hat{\mu}_a(t), q\right) \leq \log\left(\frac{n}{N_a(t)}\right)\right\}$$

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

$$I_a(t) = \max\left\{q \geq \hat{\mu}_a(t), N_a(t) d\left(\hat{\mu}_a(t), q\right) \leq \log\left(\frac{T}{N_a(t)}\right)\right\}$$

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

Approximating the FH-Gittins indices

- [Burnetas and Katehakis, 03]: when n is large,

$$G(\pi_a^t, n) \simeq \max\left\{q \geq \hat{\mu}_a(t), N_a(t) d\left(\hat{\mu}_a(t), q\right) \leq \log\left(\frac{n}{N_a(t)}\right)\right\}$$

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

$$I_a(t) = \max\left\{q \geq \hat{\mu}_a(t), N_a(t) d\left(\hat{\mu}_a(t), q\right) \leq \log\left(\frac{T}{N_a(t)}\right)\right\}$$

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

ASYMPTOTIC OPTIMALITY ?

Outline

The Bayes-UCB algorithm

π_a^t the posterior distribution over μ_a at the end of round t.

Algorithm: Bayes-UCB [K., Cappé, Garivier 2012]

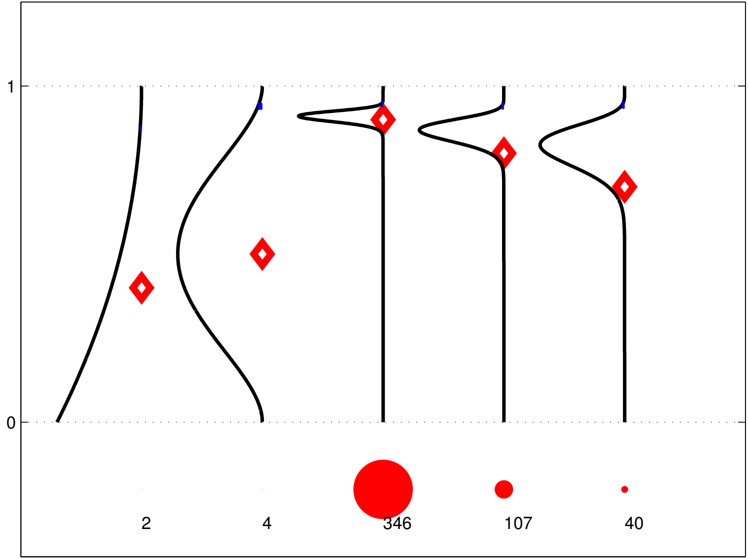
$$A_{t+1} = \underset{a}{\text{argmax}}\ Q\left(1 - \frac{1}{t(\log t)^c}, \pi_a^t\right)$$

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

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 in practice



Theory

$\nu^{\mu_1}, \dots, \nu^{\mu_K}$ are such that $\mu_a \in \mathcal{J}$ ($\mathcal{J}$ open interval)

Assumptions

$\pi = \pi_1^0 \otimes \cdots \otimes \pi_K^0$ is such that

- π_a^0 has a density h_a with respect to the Lebesgue measure
- $\forall u \in \mathcal{J}, \ h_a(u) > 0$

- The posterior distribution depends on two sufficient statistics:

$$\pi_a^t = \pi_{a,N_a(t),\hat{\mu}_a(t)}$$

An important rewriting of the posterior

$$\pi_{a,n,x}(\mathcal{I}) = \frac{\int_{\mathcal{I}} e^{-nd(x,u)} h_a(u) du}{\int_{\mathcal{J}} e^{-nd(x,u)} h_a(u) du}.$$

Theory

- Bayes-UCB rewrites

$$A_{t+1} = \underset{a}{\operatorname{argmax}} \ Q\left(1 - \frac{1}{t(\log t)^c}, \pi_{a,N_a(t),\hat{\mu}_a(t)}\right)$$

Extra assumption

Bounds on the means of the arms are known: there exists μ^-, μ^+ in $\mathcal{J}$ such that for all a, $\mu_a \in [\mu^-, \mu^+]$

Theorem

Let $\overline{\mu}_a(t) = (\hat{\mu}_a(t) \vee \mu^-) \wedge \mu^+$. The index policy

$$A_{t+1} = \underset{a}{\operatorname{argmax}} \ Q\left(1 - \frac{1}{t(\log t)^c}, \pi_{a,N_a(t),\overline{\mu}_a(t)}\right)$$

with parameter $c \geq 7$ is such that, for all $\epsilon > 0$,

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

A key element: Posterior bounds

Recall that $\pi_{a,n,x}(\mathcal{I}) = \frac{\int_{\mathcal{I}} e^{-nd(x,u)} h_a(u) du}{\int_{\mathcal{J}} e^{-nd(x,u)} h_a(u) du}$.

Bounds on the tail of the posterior distribution

The exists constants A, B, C such that, for all a, for all $n \in \mathbb{N}^*$ and $(x, v) \in [\mu^-, \mu^+]^2$,

1. if $v > x$, $An^{-1} e^{-nd(x,v)} \leq \pi_{a,n,x}([v, \mu^+[) \leq B\sqrt{n} e^{-nd(x,v)}$
2. if $v < x$, $\pi_{a,n,x}([v, \mu^+[) \geq 1/(C\sqrt{n} + 1)$

A key element: Posterior bounds

1. if $v > x$, $An^{-1}e^{-nd(x,v)} \leq \pi_{a,n,x}([v, \mu^+[) \leq B\sqrt{n}e^{-nd(x,v)}$
2. if $v < x$, $\pi_{a,n,x}([v, \mu^+[) \geq 1/(C\sqrt{n} + 1)$

Example of use:

$$
\begin{aligned}
\{\mu_1 \geq \overline{q}_1(t)\} &= \left\{\pi_{1,N_1(t),\overline{\mu}_1(t)}([\mu_1, \mu^+[) \leq \frac{1}{t\log^c t}\right\} \\
&\subset \left\{\frac{1}{C\sqrt{N_1(t)} + 1} \leq \frac{1}{t\log^c t}\right\} \bigcup \left\{\frac{Ae^{-N_1(t)d^+(\overline{\mu}_1(t),\mu_1)}}{N_1(t)} \leq \frac{1}{t\log^c t}\right\}, \\
&\subset \left\{N_1(t)d^+(\hat{\mu}_1(t), \mu_1) \geq \log\left(\frac{At\log^c t}{N_1(t)}\right)\right\},
\end{aligned}
$$

for t large enough.

An interesting by-product of our analysis

- We managed to handle alternative exploration rates !

Index policy: KL-UCB-H$^+$

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

Index policy: KL-UCB$^+$

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

The index policy associated to the indices $u_a^{H,+}(t)$ and $u_a^{+}(t)$ satisfy, for all $\epsilon > 0$,

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

Outline

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"

- the first bandit algorithm, proposed by [Thompson 1933]
- good empirical performance in complex model
- first logarithmic regret bound in 2012

Thompson Sampling is asymptotically optimal

Theorem [K.,Korda,Munos 2012],[Korda, K., Munos 2014]

For all $\epsilon > 0$,

$$\mathbb{E}[N_a(T)] \leq (1 + \epsilon) \frac{1}{d(\mu_a, \mu^*)} \log(T) + o_{\mu,\epsilon}(\log(T)).$$

This results holds:

- for Bernoulli bandits, with a uniform prior
- for exponential family bandits, with the Jeffrey's prior

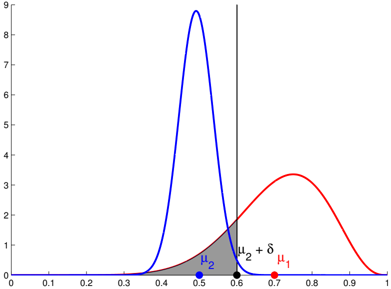
A key ingredient in the proof

Proposition

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

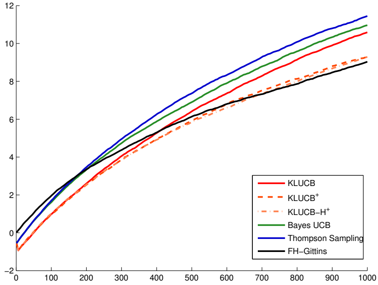
$$\sum_{t=1}^{\infty} \mathbb{P}\left(N_1(t) \leq t^b\right) \leq C_b.$$

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

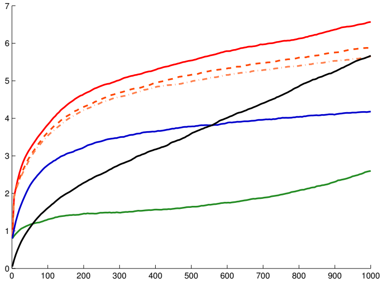


Numerical experiments

- Short horizon, $T = 1000$ (average over $N = 10000$ runs)



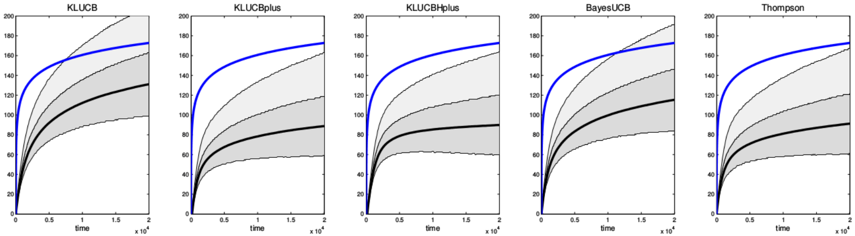
$\mu_1 = 0.2, \mu_2 = 0.25$



$\mu_1 = 0.85, \mu_2 = 0.95$

Numerical experiments

- Long horizon, $T = 20000$ (average over $N = 50000$ runs)



10 arms bandit problem

$$\mu = [0.1\ 0.05\ 0.05\ 0.05\ 0.02\ 0.02\ 0.02\ 0.01\ 0.01\ 0.01]$$

Outline

Contextual linear bandit models

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

$=$ characteristics of the items to recommend

The model:

- if the context $x_t \in \mathcal{D}_t$ is selected
- a reward $r_t = x_t^T \theta + \epsilon_t$ is received

$\theta =$ underlying preference vector

Bayesian model: (with Gaussian prior)

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

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

$$\left\{ \begin{array}{lcl} \hat{\theta}(t) & = & (B(t))^{-1} X_t^T Y_t \ \text{ with } \ B(t) = \frac{\sigma^2}{\kappa^2} I_d + \sum_{s=1}^t x_s x_s^T \\ \Sigma_t & = & \sigma^2 (B(t))^{-1}. \end{array} \right.$$

Contextual linear bandit models

- **Bayes-UCB**

$$
\begin{aligned}
x_{t+1} &= \underset{x \in \mathcal{D}_{t+1}}{\operatorname{argmax}} \; Q(1 - e^{-f(t)}, \mathcal{L}(x^T \theta | x_1, r_1, \dots, x_t, r_t)) \\
&= \underset{x \in \mathcal{D}_{t+1}}{\operatorname{argmax}} \; x^T \hat{\theta}(t) + ||x||_{\Sigma_t} Q(1 - e^{-f(t)}, \mathcal{N}(0,1))
\end{aligned}
$$

- **Thompson Sampling**

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

Theoretical guarantees

Bayesian guarantees: ($|\mathcal{D}_t| \leq K$)

$$\mathbb{E}_{\theta\sim\mathcal{N}(0,\kappa^2)}\left[\sum_{t=1}^{T}\left(\max_{x\in\mathcal{D}_t} x^T\theta - x_t^T\theta\right)\right] = O_{\kappa^2,\sigma^2}\left(\sqrt{d\,T\log(K)}\right)$$

for Bayes-UCB [K. 2014], TS [Russo, Van Roy 2013]

Frequentist guarantees: [Agrawal, Goyal, 2013]

With $\kappa = v = \sigma\sqrt{9d\log(T^2)}$, for TS based on the model

$$\theta \sim \mathcal{N}(0,\kappa^2), \quad \epsilon_t \sim \mathcal{N}(0,v^2),$$

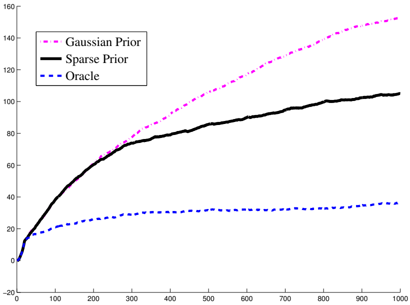
$$\mathbb{E}_{\theta}\left[\sum_{t=1}^{T}\left(\max_{x\in\mathcal{D}_t} x^T\theta - x_t^T\theta\right)\right] = O_{\kappa^2,\sigma^2}\left(d\sqrt{T\log(K)}\right)$$

Open questions: choice of prior? optimal dependency in d?

Beyond Gaussian prior...

- A sparsity-inducing prior (spike-and-slab)

$$\forall a = 1, \dots, K, \quad \theta_a \sim \epsilon\delta_0 + (1-\epsilon)\mathcal{N}(0, \kappa^2) \ .$$

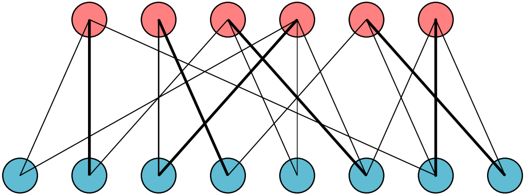


Bayes-UCB, $\quad d = 10, \quad K = 20, \quad \theta = [1, 1, 0, \dots, 0]$

Wanted: Good MCMC sampler for experiments in large dimension

Thompson Sampling in combinatorial bandit models

- Arms are edges on a graph
- $\mathcal{M}$ is a set of possible configurations (subsets of edges)
- The agent chooses $m_t \in \mathcal{M}$ at time t and observe a realization of all arms in m_t (semi-bandit)
- Goal: play as much as possible the best configuration $m^* \in \mathcal{M}$



TS: sample the weights on all edges from a posterior distribution, choose the best configuration in this sampled weighted graph

Conclusion

Several index policies inspired by the Bayesian MAB:

- FH-Gittins, based on the finite-horizon Gittins indices
- KL-UCB$^+$and KL-UCB-H$^+$, two variants of KL-UCB using an alternative exploration rate, inspired by the Bayesian solution
- Bayes-UCB, based on posterior quantiles
- Thompson Sampling, based on posterior samples

... evaluated in terms of (frequentist) regret:

- good empirical performance
- (almost) all are asymptotically optimal in simple models

Bayes-UCB and TS are easier to implement than KL-UCB in simple models, and can be easily used in more complex models