Analyse de stratégies bayésiennes et fréquentistes pour l'allocation séquentielle de ressources

Emilie Kaufmann

Sequential allocation: some examples

Clinical trial

- K possible treatments (with unknown effect)

- Which treatment should be allocated to each patient based on their effect on previous patients?

Movie recommendation

- K different movies

- Which movie should be recommended to each user, based on the ratings given by previous (similar) users?

The 'bandit' framework

One-armed bandit
= slot machine (or arm)

<u>Multi-armed bandit:</u> several arms.
Drawing arm $a \Leftrightarrow$ observing a sample from a distribution ν_a, with mean μ_a

Best arm $a^* = \text{argmax}_a \; \mu_a$

Which arm should be drawn based on the previous observed outcomes?

Outline

Two bandit problems

Regret minimization: a Bayesian approach

Best arm identification: towards optimal algorithms

Bandit model

A **multi-armed bandit model** is a set of K arms where

- Each arm a is a probability distribution ν_a of mean μ_a
- Drawing arm a is observing a realization of ν_a
- Arms are assumed to be independent

At round t, an agent

- chooses arm A_t, and observes $X_t \sim \nu_{A_t}$
- (A_t) is his **strategy** or **bandit algorithm**, such that

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

Global objective: Learn which arm(s) have highest mean(s)

$$\mu^* = \max_a \mu_a \qquad a^* = \text{argmax}_a \, \mu_a$$

Rewards maximization or regret minimization

Samples are seen as **rewards**.

The agent ajusts (A_t) to

- maximize the (expected) sum of rewards accumulated,

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

- or equivalently minimize his *regret*:

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

$\Rightarrow$ **exploration/exploitation tradeoff**

Best arm identification

The agent has to **identify the set of m best arms $\mathcal{S}_m^*$**
(no loss when drawing 'bad' arms)

He

- uses a sampling strategy (A_t)
- stops at some (random) time τ
- recommends a subset $\hat{S}$ of m arms

His goal:

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

$\Rightarrow$ **optimal exploration**

Back to the example of medical trials

K possible treatments for a given symptom.

- treatment number a has (unknown) probability of success μ_a

The doctor:

- chooses treatment A_t to give to patient t
- observes whether the patient is cured : $X_t \sim \mathcal{B}(\mu_{A_t})$

He can ajust his strategy (A_t) so as to

Regret minimization	Best arm identification
Maximize the number of patients cured among T patients	Identify the best treatment with probability at least $1-\delta$ (to always give this one later)

Questions

- Are Bayesian algorithms efficient when evaluated with (frequentist) regret?
- Can recent improvements for regret minimization be transposed to the best arm identification framework?
- What is an *optimal* algorithm for best arm identification?

Two probabilistic modelings

K independent arms.

Frequentist model	**Bayesian model**
$\theta_1, \dots, \theta_K$ unknown parameters	$\theta_1, \dots, \theta_K$ drawn from a prior distribution : $\theta_a \overset{i.i.d.}{\sim} \pi_a$
$(X_{a,t})_t \overset{\text{i.i.d.}}{\sim} \nu_{\theta_a}$	$(X_{a,t})_t \| \theta_a \overset{\text{i.i.d.}}{\sim} \nu_{\theta_a}$

At time t, arm A_t is drawn and $X_t = X_{A_t,t}$.

Two measures of performance

Regret	**Bayes risk**
$R_T(\theta) = \mathbb{E}_\theta \left[\sum_{t=1}^T (\mu^* - \mu_{A_t}) \right]$	$\text{Risk}_T(\pi) = \mathbb{E}\left[\sum_{t=1}^T (\mu^* - \mu_{A_t}) \right] = \int R_T(\theta) d\pi(\theta)$

Frequentist tools, Bayesian tools

Bandit algorithms based on frequentist tools use:

- Maximum Likelihood Estimator of the mean of each arm
- Confidence Intervals for the mean of each arm

Bandit algorithms based on Bayesian tools use:

- $\Pi_t = (\pi_1^t, \ldots, \pi_K^t)$ the current posterior over $(\theta_1, ..., \theta_K)$

One can **separate tools and objectives**:

Performance criterion	Frequentist algorithms	Bayesian algorithms
Regret	?	?
Bayes risk	?	?

Our goal: propose Bayesian algorithms optimal w.r.t. the regret

Optimal algorithms for regret minimization

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

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

- [Lai and Robbins 1985]: every consistent policy satisfies

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

Definition

A bandit algorithm is **asymptotically optimal** if, for every θ,

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

Towards asymptotically optimal algorithms

- A UCB-type algorithm chooses at time $t+1$

$$A_{t+1} = \underset{a}{\arg\max}\ UCB_a(t)$$

where $UCB_a(t)$ is some upper confidence bound. 2

Examples for binary bandits (Bernoulli distributions)

- UCB1 [Auer et al. 02] uses Hoeffding bounds:

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

$S_a(t)$: sum of rewards from arm a up to time t

$$\mathbb{E}[N_a(T)] \leq \frac{K_1}{2(\mu_a - \mu^*)^2} \log T + K_2, \quad \text{with } K_1 > 1.$$

KL-UCB: an asymptotically optimal algorithm

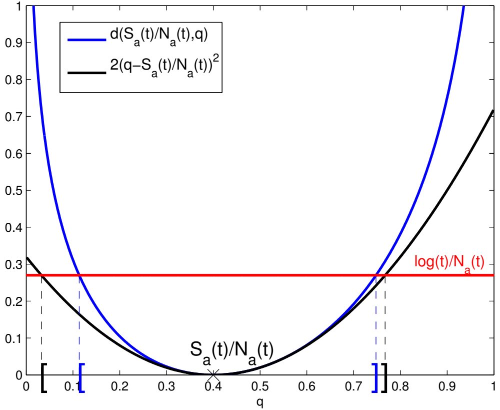
- KL-UCB [Cappé et al. 2013] uses the index:

$$u_a(t) = \underset{x > \frac{S_a(t)}{N_a(t)}}{\operatorname{argmax}} \left\{ d\left(\frac{S_a(t)}{N_a(t)}, x\right) \leq \frac{\log(t) + c \log\log(t)}{N_a(t)} \right\},$$

where

$$\begin{aligned} d(p,q) &= \mathrm{KL}\left(\mathcal{B}(p), \mathcal{B}(q)\right) \\ &= p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right). \end{aligned}$$

KL-UCB: an asymptotically optimal algorithm



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

Optimal algorithms for Bayes risk minimization

There exists an exact solution to Bayes risk minimization, that satisfies dynamic programming equations.

Bernoulli bandit model $\nu = (\mathcal{B}(\theta_1), \dots, \mathcal{B}(\theta_K))$

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

The history of the game up to time t can be summarized by a posterior matrix $\mathcal{S}_t$

- $\mathcal{S}_t$ can be seen as a state in a **Markov Decision Process**.

Optimal algorithms for Bayes risk minimization

There exists an optimal policy (A_t) in this MDP satisfying

$$\underset{(A_t)}{\arg\max}\ \mathbb{E}\left[\sum_{t=1}^{\infty} \alpha^{t-1} X_t\right] \quad \text{or} \quad \underset{(A_t)}{\arg\max}\ \mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$$

- [Gittins'79]: in the discounted case, the solution reduces to an index policy (Gittins indices)
- with a finite horizon, this reduction no longer holds

However:

- FH-Gittins, the index policy associated to Finite-Horizon Gittins indices, performs well in practice (even w.r.t. regret), but its implementation is costly

Summary so far

Objective	Frequentist algorithms	Bayesian algorithms
Regret	KL-UCB	?
Bayes risk	KL-UCB-H^+ [Lai 87]	Dynamic Programming FH-Gittins ?

UCBs versus Bayesian algorithms

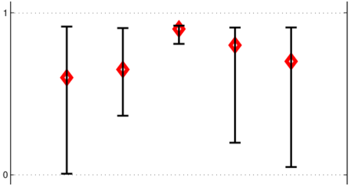


Figure: Confidence intervals on the means of the arms after t rounds

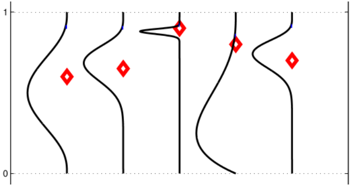


Figure: Posterior distribution of the means of the arms after t rounds

$\Rightarrow$ How do we exploit the posterior in a Bayesian bandit algorithm?

The Bayes-UCB algorithm

Let :

- $\Pi_0 = (\pi_1^0, \dots, \pi_K^0)$ be a prior distribution over $(\theta_1, ..., \theta_K)$
- $\Lambda_t = (\lambda_1^t, \dots, \lambda_K^t)$ be the posterior over the means $(\mu_1, ..., \mu_K)$ a the end of round t

Algorithm: Bayes-UCB

The **Bayes-UCB algorithm** chooses at time $t+1$

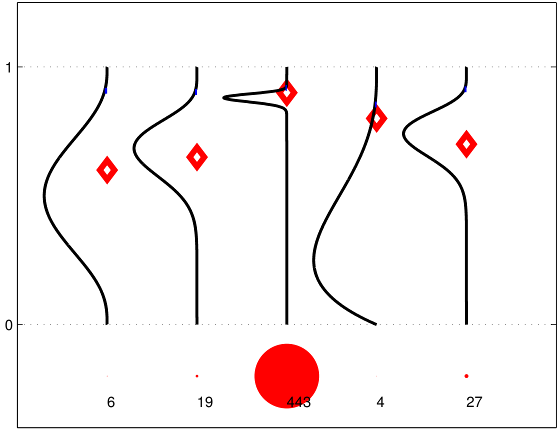
$$A_{t+1} = \underset{a}{\text{argmax}}\ Q\left(1 - \frac{1}{t(\log t)^c}, \lambda_a^{t-1}\right)$$

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

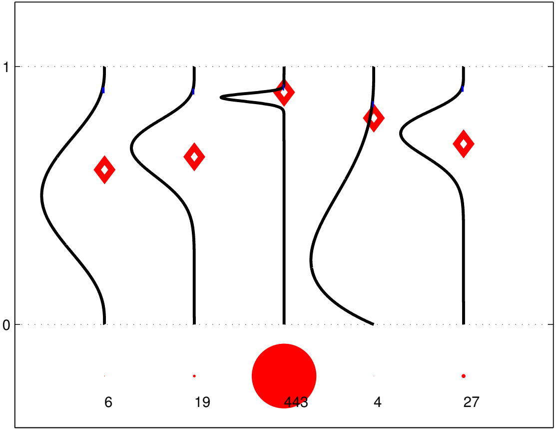
Bernoulli reward with uniform prior: $\theta = \mu$ and $\Pi_t = \Lambda_t$

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

Bayes-UCB in action !



Bayes-UCB in action !



Theoretical results in the Bernoulli case

- **Bayes-UCB is asymptotically optimal**

Theorem [K.,Cappé,Garivier 2012]

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

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

Links to a frequentist algorithm

Bayes-UCB index is close to KL-UCB indices:

Lemma

$$\tilde{u}_a(t) \leq q_a(t) \leq u_a(t)$$

with:

$$u_a(t) = \underset{x > \frac{S_a(t)}{N_a(t)}}{\text{argmax}} \left\{ d\left(\frac{S_a(t)}{N_a(t)}, x\right) \leq \frac{\log(t) + c \log\log(t)}{N_a(t)} \right\}$$

$$\tilde{u}_a(t) = \underset{x > \frac{S_a(t)}{N_a(t)+1}}{\text{argmax}} \left\{ d\left(\frac{S_a(t)}{N_a(t)+1}, x\right) \leq \frac{\log\left(\frac{t}{N_a(t)+2}\right) + c \log\log(t)}{(N_a(t)+1)} \right\}$$

Bayes-UCB appears to build automatically confidence intervals based on Kullback-Leibler divergence, that are adapted to the geometry of the problem.

Ingredients of the proof

We have tight bounds on the tail of posterior distributions (Beta distributions)

- First element: link between Beta and Binomial distribution:

$$\mathbb{P}(X_{a,b} \geq x) = \mathbb{P}(S_{a+b-1,1-x} \geq b)$$

- Second element: Sanov inequality: for $k > nx$,

$$\frac{e^{-nd\left(\frac{k}{n},x\right)}}{n+1} \leq \mathbb{P}(S_{n,x} \geq k) \leq e^{-nd\left(\frac{k}{n},x\right)}$$

Thompson Sampling

$\Pi^t = (\pi_1^t, .., \pi_K^t)$ posterior distribution on $(\theta_1, .., \theta_K)$ at round t.

$\mu(\theta)$ the mean of an arm arm parametrized by θ.

Algorithm: Thompson Sampling

Thompson Sampling is a randomized Bayesian algorithm:

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

General principle: Each arm is drawn according to its posterior probability of being optimal

- the first bandit algorithm, proposed in 1933 [Thompson 1933]
- his good empirical performances are demostrated beyond the Bernoulli case [Scott, 2010],[Chapelle, Li 2011]
- no regret upper bound before 2012...

An optimal regret bound for Bernoulli bandits

- A first result: [Agrawal,Goyal 2012]

$$R_T(\theta) \leq C \left(\sum_{a \neq a^*} \frac{1}{(\mu^* - \mu_a)^2} \right)^2 \log(T) + o_\mu(\log(T))$$

- Our improvement:

Theorem [K.,Korda,Munos 2012]

For all $\epsilon > 0$,

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

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

Ingredients of the proof

Let arm a be suboptimal and arm 1 be the optimal arm.

- A new decomposition

$$(A_{t+1} = a) \subseteq$$

$$\left(\theta_1(t) \leq \mu_1 - \sqrt{\frac{6 \log t}{N_1(t)}}\right) \bigcup \left(\theta_a(t) \geq \mu_1 - \sqrt{\frac{6 \log t}{N_1(t)}}, A_{t+1} = a\right)$$

- Prove that

$$\sum_{t=0}^{\infty} \mathbb{P}\left(\theta_1(t) \leq \mu_1 - \sqrt{\frac{6 \log t}{N_1(t)}}\right) < +\infty.$$

- Use a quantile to replace the sample:

$$q_a(t) := Q\left(1 - \frac{1}{t \log(T)}, \pi_a^t\right) \Rightarrow \sum_{t=1}^{T} \mathbb{P}\left(\theta_a(t) > q_a(t)\right) \leq 2$$

and use what we know about quantiles (cf. Bayes-UCB)

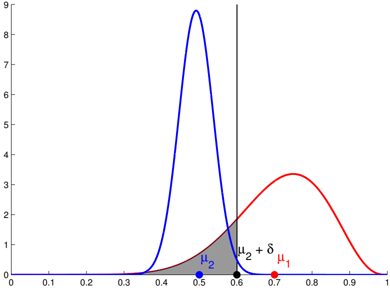
A key ingredient

Proposition

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

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

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



Summary: our contributions to regret minimization

Objective	Frequentist algorithms	Bayesian algorithms
Regret	KL-UCB	Bayes-UCB Thompson Sampling
Bayes risk	KL-UCB-H$^+$	Dynamic Programming FH-Gittins?

Bayes-UCB and Thompson Sampling are good alternatives to KL-UCB, asymptotically optimal in the Bernoulli case and easy to implement, even in more complex models.

Other contributions:

- analysis of Thompson Sampling for rewards in a one-parameter exponential family
- Bayes risk bounds for Bayes-UCB and Thompson Sampling for contextual linear bandit problems (Chapter 4)

Two bandit problems

Regret minimization: a Bayesian approach

Bayes-UCB

Thompson Sampling

Best arm identification: towards optimal algorithms

m best arms identification in the fixed-confidence setting

Assume $\mu_1 \geq \cdots \geq \mu_m > \mu_{m+1} \geq \ldots \mu_K$ (Bernoulli bandit model)

Parameters and notations

- m a fixed number of arms to find
- $\delta \in]0, 1[$ a risk parameter
- $\mathcal{S}_m^* = \{1, \ldots, m\}$ the set of m optimal arms

The agent:

- draws arm A_t at time t
- decides to stop after a (possibly random) total number of samples from the arms τ
- recommends a set $\hat{\mathcal{S}}$ of m arms

His goal:

- the algorithm is δ-PAC : $\forall \nu \in \mathcal{M}, \mathbb{P}_\nu(\hat{\mathcal{S}} \neq \mathcal{S}_m^*) \leq \delta$.
- the sample complexity $\mathbb{E}_\nu[\tau]$ is small

The complexity of best-arm identification

The literature presents δ-PAC algorithm such that

$$\mathbb{E}_\nu[\tau] \leq C H(\nu) \log(1/\delta)$$

[Mannor Tsitsiklis 04],[Even-Dar et al. 06], [Kalyanakrishnan et al.12]

In order to compute the complexity term

$$\boxed{\inf_{\substack{\delta-\text{PAC} \\ \text{algorithms}}} \limsup_{\delta \to 0} \frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)}}$$

we need:

- ➜ a lower bound on $\mathbb{E}_\nu[\tau]$
- ➜ an algorithm reaching the lower bound

Lower bound: changes of distribution

- A new formulation for a change of distribution:

Lemma [K., Cappé, Garivier 2014]

$\nu = (\nu_1, \nu_2, \dots, \nu_K)$, $\nu' = (\nu'_1, \nu'_2, \dots, \nu'_K)$ two bandit models, A an event,

$$\sum_{a=1}^{K} \mathbb{E}_\nu[N_a(\tau)] \mathrm{KL}(\nu_a, \nu'_a) \geq d(\mathbb{P}_\nu(A), \mathbb{P}_{\nu'}(A)).$$

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

Apply it to

- ν and ν' such that $\mathcal{S}^*_m(\nu) \neq \mathcal{S}^*_m(\nu')$
- $A = (\hat{S} = \mathcal{S}^*_m(\nu))$: $\mathbb{P}_\nu(A) \geq 1 - \delta$ and $\mathbb{P}_{\nu'}(A) \leq \delta$

Lower bound: a general result

Theorem [K.,Cappé, Garivier 14]

Any algorithm that is δ-PAC on every binary bandit model such that $\mu_m > \mu_{m+1}$ satisfies, for $\delta \leq 0.15$,

$$\mathbb{E}[\tau] \geq \left(\sum_{a=1}^{m} \frac{1}{d(\mu_a, \mu_{m+1})} + \sum_{a=m+1}^{K} \frac{1}{d(\mu_a, \mu_m)} \right) \log \frac{1}{2\delta}$$

- First lower bound for $m > 1$
- Involves information-theoretic quantities

An algorithm: KL-LUCB

Generic notation:

- confidence interval (C.I.) on the mean of arm a at round t:

$$\mathcal{I}_a(t) = [L_a(t), U_a(t)]$$

- $J(t)$ the set of m arms with highest empirical means

Our contribution: Introduce KL-based confidence intervals

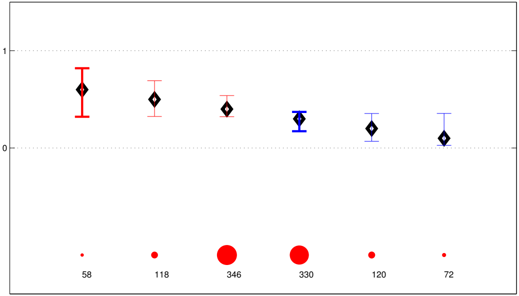
$$\begin{aligned} U_a(t) &= \max\left\{q \geq \hat{\mu}_a(t) : N_a(t) d(\hat{\mu}_a(t), q) \leq \beta(t,\delta)\right\} \\ L_a(t) &= \min\left\{q \leq \hat{\mu}_a(t) : N_a(t) d(\hat{\mu}_a(t), q) \leq \beta(t,\delta)\right\} \end{aligned}$$

for $\beta(t, \delta)$ some exploration rate.

An algorithm: KL-LUCB

At round t, the algorithm:

- draws two well-chosen arms: u_t and l_t (in bold)
- stops when C.I. for arms in $J(t)$ and $J(t)^c$ are separated



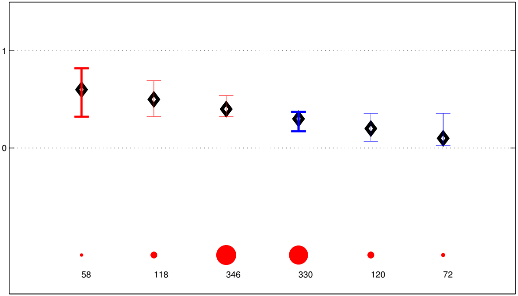
$$m = 3, K = 6$$

Set $J(t)$, arm l_t in bold Set $J(t)^c$, arm u_t in bold

An algorithm: KL-LUCB

At round t, the algorithm:

- draws two well-chosen arms: u_t and l_t (in bold)
- stops when C.I. for arms in $J(t)$ and $J(t)^c$ are separated



$$m = 3, K = 6$$

Set $J(t)$, arm l_t in bold Set $J(t)^c$, arm u_t in bold

Theoretical guarantees

Theorem [K.,Kalyanakrishnan 2013]

KL-LUCB using the exploration rate

$$\beta(t,\delta) = \log\left(\frac{k_1 K t^{\alpha}}{\delta}\right),$$

with $\alpha > 1$ and $k_1 > 1 + \frac{1}{\alpha - 1}$ satisfies $\mathbb{P}(\hat{\mathcal{S}} = \mathcal{S}_m^*) \geq 1 - \delta$.
For $\alpha > 2$,

$$\mathbb{E}[\tau] \leq 4\alpha H^* \log\left(\frac{1}{\delta}\right) + o_{\delta \to 0}\left(\log \frac{1}{\delta}\right),$$

with

$$H^* = \min_{c \in [\mu_{m+1};\mu_m]} \sum_{a=1}^{K} \frac{1}{d^*(\mu_a, c)}.$$

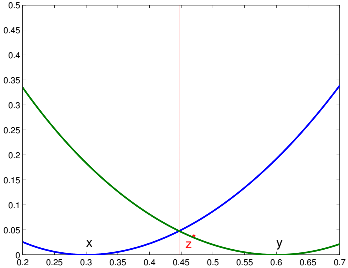
Theoretical guarantees

- **Another informational quantity: Chernoff information**

$$d^*(x, y) := d(z^*, x) = d(z^*, y),$$

where z^* is defined by the equality

$$d(z^*, x) = d(z^*, y).$$



Bounds on the complexity term

Lower bound

$$\inf_{\substack{\delta-\text{PAC}\\ \text{algorithms}}} \limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau]}{\log\frac{1}{\delta}} \geq \sum_{t=1}^{m} \frac{1}{d(\mu_a, \mu_{m+1})} + \sum_{t=m+1}^{K} \frac{1}{d(\mu_a, \mu_m)}$$

Upper bound (for KL-LUCB)

$$\inf_{\substack{\delta-\text{PAC}\\ \text{algorithms}}} \limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau]}{\log\frac{1}{\delta}} \leq 8 \min_{c\in[\mu_{m+1};\mu_m]} \sum_{a=1}^{K} \frac{1}{d^*(\mu_a, c)}$$

Summary: our contributions to best arm identification

We proposed:

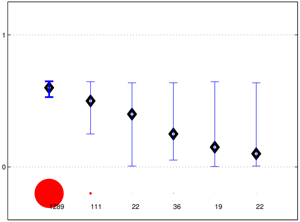
- a new lower bound for $m > 1$
- the analysis of KL-LUCB, that successfully transposes recent improvements from the regret minimization to the best arm identification framework

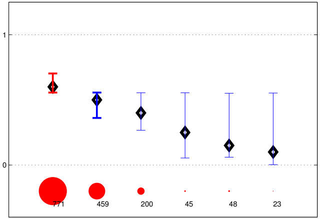
Other contributions: (see Chapter 5)

- refined lower bounds in the two-armed case in the fixed-confidence and fixed-budget settings
- characterization of the complexity of both settings for some classes of two-armed bandit models

Conclusion and perspectives

- KL-based confidence intervals are useful for both regret minimization and best arm identification





- Bayesian algorithms are efficient (and optimal) for solving the (frequentist) regret minimization problem

Some remaining questions:

- What information-theoretic quantity characterizes the complexity of best arm identification when $m > 1$?
- Can Bayesian tools be used for best arm identification as well?

Two complexity terms

Let $\mathcal{M}$ be a class of bandit models.
An algorithm $\mathcal{A} = ((A_t), \tau, \hat{S})$ is...

Fixed-confidence setting	Fixed-budget setting
δ-PAC on $\mathcal{M}$ if $\forall \nu \in \mathcal{M}$, $\mathbb{P}_\nu(\hat{\mathcal{S}} \neq \mathcal{S}_m^*) \leq \delta$	consistent on $\mathcal{M}$ if $\forall \nu \in \mathcal{M}$, $p_t(\nu) := \mathbb{P}_\nu(\hat{S}_t \neq \mathcal{S}_m^*) \underset{t\to\infty}{\longrightarrow} 0$

Two complexities

$\kappa_{\mathrm{C}}(\nu) = \inf_{\mathcal{A}\ \delta-\mathrm{PAC}} \limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)}$	$\kappa_{\mathrm{B}}(\nu) = \inf_{\mathcal{A}\ \mathrm{cons.}} \left(\limsup_{t\to\infty} -\frac{1}{t}\log p_t(\nu)\right)^{-1}$
for a probability of error $\leq \delta$ $\mathbb{E}_\nu[\tau] \simeq \kappa_C(\nu) \log\frac{1}{\delta}$	for a probability of error $\leq \delta$, budget $t \simeq \kappa_B(\nu) \log\frac{1}{\delta}$

The complexity of A/B Testing

- Refined lower bounds for two-armed bandits

Fixed-confidence setting	Fixed-budget setting
any δ-PAC algorithm satisfies $\mathbb{E}_\nu[\tau] \geq \frac{1}{d_*(\mu_1,\mu_2)} \log\left(\frac{1}{2\delta}\right)$	any consistent algorithm satisfies $\limsup_{t\to\infty} -\frac{1}{t}\log p_t(\nu) \leq d^*(\mu_1,\mu_2)$

where

$$d_*(x,y) := d(x,z_*) = d(y,z_*),$$

with z_* defined by

$$d(x,z_*) = d(y,z_*).$$

$$\boxed{d^*(\mu_1,\mu_2) > d_*(\mu_1,\mu_2)}$$

The complexity of A/B Testing

- In the fixed-budget setting, for every ν, there exists an algorithm such that

$$\limsup_{t\to\infty} -\frac{1}{t}\log p_t(\nu) \geq d^*(\mu_1,\mu_2)$$

Thus for Bernoulli bandit models

$$\boxed{\kappa_B(\nu) = \frac{1}{d^*(\mu_1,\mu_2)}} \quad \text{and} \quad \boxed{\kappa_B(\nu) \geq \kappa_C(\nu)}$$

- For two-armed Gaussian bandit models

$$\mathcal{M} = \left\{\nu = \left(\mathcal{N}\left(\mu_1,\sigma_1^2\right), \mathcal{N}\left(\mu_2,\sigma_2^2\right)\right) : (\mu_1,\mu_2) \in \mathbb{R}^2, \mu_1 \neq \mu_2\right\},$$

$$\boxed{\kappa_B(\nu) = \kappa_C(\nu) = \frac{2(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2}}$$

Contextual linear bandit models

At time t:

- a set of 'contexts' $\mathcal{D}_t \subset \mathbb{R}^d$ is revealed
- the agent chooses $x_t \in \mathcal{D}_t$
- he receives a reward

$$y_t = x_t^T \theta + \epsilon_t.$$

His goal: minimizing

$$\mathcal{R}_\theta(T, \mathcal{A}) = \sum_{t=1}^{T} (x_t^*)^T \theta - x_t^T \theta$$

where

$$x_t^* = \underset{x \in \mathcal{D}_t}{\arg\max}\, x^T \theta.$$

Bayes-UCB and Thompson Sampling

Bayesian model:

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

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

- **Bayes-UCB**

$$\begin{aligned} x_{t+1} &= \underset{x \in \mathcal{D}_{t+1}}{\text{argmax}}\ Q\left(1 - e^{-f(t+1,\delta)}; \mathcal{N}\left(x^T\hat{\theta}(t), ||x||_{\Sigma_t}\right)\right), \\ x_{t+1} &= \underset{x \in \mathcal{D}_{t+1}}{\text{argmax}}\ \left[x^T\hat{\theta}(t) + ||x||_{\Sigma_t} Q\left(1 - e^{-f(t+1,\delta)}; \mathcal{N}(0,1)\right)\right]. \end{aligned}$$

Theorem

For $f(t,\delta) = \log(\pi^2 K T^2 / 3\delta)$, if $|\mathcal{D}_t| = K$ for all t,

$$\mathbb{P}\left(\mathcal{R}_\theta(T, \mathcal{A}) = \tilde{O}\left(\sqrt{dT\log(K)}\right)\right) \geq 1 - \delta.$$

Bayes-UCB and Thompson Sampling

Bayesian model:

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

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

- **Thompson Sampling**

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

Theorem

Without any assumption on the number of contexts in $\mathcal{D}_t$,

$$\mathbb{E}\left[\mathcal{R}_\theta(T, \mathrm{TS})\right] = \tilde{O}\left(d\sqrt{T}\right).$$