Multi-armed bandits and boundary crossing probabilites

November 06, Sapporo

Odalric-Ambrym Maillard, Inria Lille

SequeL
–**Seque**ntial **L**earning–
Inria - Cristal (Dating)

Philippe **Preux** (head), Emilie **Kaufmann**, Michal **Valko**, Me.

Introduction

Stochastic multi-armed bandits

Introduction

Multi-armed bandits

Regret lower-bounds

Near-optimal strategies

Boundary crossing for regret analysis

Stochastic multi-armed bandits

Sources of i.i.d. $\mathbb{R}$-valued observations:

$$\nu_1 \quad \nu_2 \quad \dots \quad \nu_{A-1} \quad \nu_A$$

Game: At each round $t \in \mathbb{N}$,

- Choose index $A_t \in \{1, \dots, A\}$
- Receive one sample $Y_t \sim \nu_{A_t}$, called the **reward**.

Goal: **maximize sum of collected rewards** $\sum_{t=1} Y_t$ over time, in expectation.

Sources are *unknown*.

- The environment ***does not*** reveal the rewards of the other arms.

Stochastic multi-armed bandits setup

- Let $\mu_\star = \max_{a\in\mathcal{A}} \mu_a$, where $\mu_a \in \mathbb{R}$ denotes the mean of ν_a.
- Let $\mathcal{A}_\star(\nu) = \mathrm{Argmax}_{a\in\mathcal{A}}\mu_a$ be the set of optimal arms.

Regret minimization

The regret captures the sub-optimality of our strategy w.r.t. an optimal one:

$$\mathfrak{R}_T \stackrel{\mathrm{def}}{=} T\mu^\star - \mathbb{E}\Big[\sum_{t=1}^T Y_t\Big] = \sum_{a\in\mathcal{A}} \underbrace{\mu_\star - \mu_a}_{\Delta_a} \mathbb{E}\Big[N_a(T)\Big].$$

$$\text{where } N_a(T) = \sum_{t=1}^T \mathbb{I}_{A_t=a}.$$

- $\mathbb{E}$ summarizes any possible source of randomness.
- Regret grows with T : we target $o(T)$ regret.

Stochastic multi-armed bandits setup

The sampling strategy (or bandit algorithm) (A_t) is sequential:

$$A_{t+1} = \pi(\underbrace{A_1, Y_1, \dots, A_t, Y_t}_{\text{past history}}).$$

- Terminology: π is the *policy* or pulling strategy.
  It may depend on past history, and be randomized.

- "i.i.d. Stochastic bandit"
  - independence between arms,
  - independence between observations of each arm (product measures),
  - stationarity (invariance by a time shift).

The learner

History at the end of round t: $H_t = (A_1, Y_1, \dots, A_t, Y_t)$.

- Learner may use H_t to base its action A_{t+1} on in round $t+1$.
- Learner uses a **"policy"**: a map π of all possible histories $\mathcal{H}$ to actions $\mathcal{A}$.
- The learner is also allowed to randomize : $\pi : \mathcal{H} \to \mathcal{P}(\mathcal{A})$, where $\mathcal{P}(\mathcal{A})$ denotes probability measures over the set $\mathcal{A}$.
- The learner may or not know the number of interaction steps with the environment.

Why do we care?

Basic model (first approximation) for:

- Clinical trials: (Thompson, 1933)

 :     

Why do we care?

Basic model (first approximation) for:

- Clinical trials: (Thompson, 1933)

:

- Casino slot machines: (Robbins, 1952)

:

Why do we care?

Basic model (first approximation) for:

- Clinical trials: (Thompson, 1933)

:




- Casino slot machines: (Robbins, 1952)

:


- Ad-placement: (Nowadays...)

:


Example of rewards

- $Y_t = 1$ if user clicks on displayed add/link/news, 0 else.
- $Y_t =$ time spent before closing a video-add.
- $Y_t =$ health status of a patient.
- ...

Design of rewards is not easy in general, and may greatly affect the behavior of an optimal agent.

Why do we care?

Building bloc for many challenging problems (+10k papers):

- Which post from your friends to show you on Facebook? (Recommender system)

Why do we care?

Building bloc for many challenging problems (+10k papers):

- Which post from your friends to show you on Facebook? (Recommender system)
- What move should be considered next when playing chess/go? (Planning)

Why do we care?

Building bloc for many challenging problems (+10k papers):

- Which post from your friends to show you on Facebook? (Recommender system)
- What move should be considered next when playing chess/go? (Planning)
- In which order should results from a search engine be presented to you? (Ranking)
- Which parameter best calibrate this microscope? (Optimization)
- What is shortest route to deliver this message? (Packet routing)

Why do we care?

Future(?) applications:

- Plant-health care:

:

Why do we care?

Future(?) applications:

- Plant-health care:

:



- Ground-health care:

:



Why do we care?

Future(?) applications:

- Plant-health care:

 :    

- Ground-health care:

 :    

- Bio-diversity/Bio-equilibrium care:

 :  

A simple strategy: "Follow the leader"

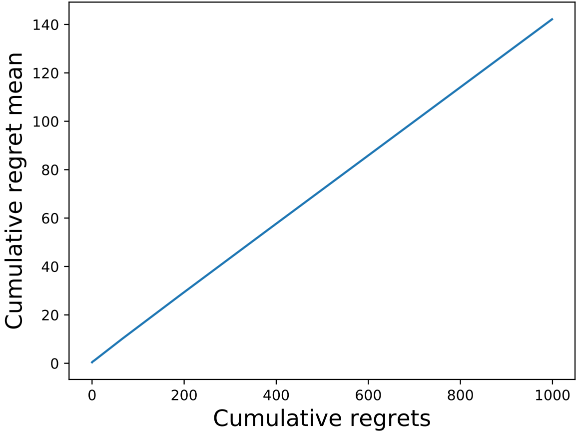
- Empirical counts: $\forall a \in \mathcal{A}$, $N_a(t) = \sum_{t'=1}^{t} \mathbb{I}\{A_{t'} = a\}$
- Empirical means: $\forall a \in \mathcal{A}$, $\tilde{\mu}_{a,t} = \frac{1}{N_a(t)} \sum_{t'=1}^{t} Y_{t'} \mathbb{I}\{A_{t'} = a\}$

$$\text{Play } A_t \in \mathrm{Argmax}_{a \in \mathcal{A}} \tilde{\mu}_{a,t}$$

- Let $\tau_{a,n} = \min\{t \geqslant 1 : N_a(t) = n\}$, $X_{a,n} = Y_{\tau_{a,n}}$, then

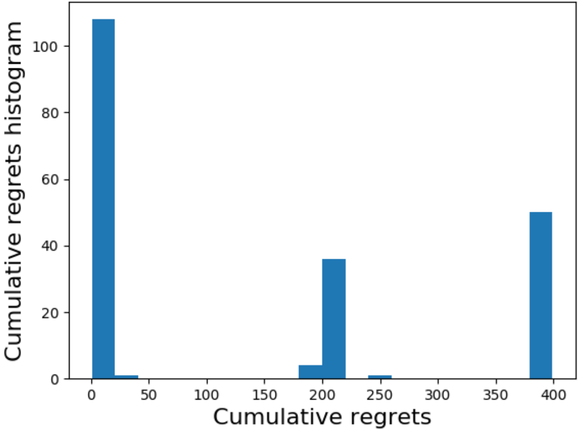
$$\tilde{\mu}_{a,t} = \widehat{\mu}_{a,N_a(t)} \text{ where } \widehat{\mu}_{a,n} = \frac{1}{n} \sum_{m=1}^{n} X_{a,m}$$

Regret on a $[\mathcal{B}(0.2), \mathcal{B}(0.4), \mathcal{B}(0.6)]$-bandit



Results averaged over 200 runs.

Regret on a $[\mathcal{B}(0.2), \mathcal{B}(0.4), \mathcal{B}(0.6)]$-bandit



A better strategy

We want to play: $\mathrm{Argmax}\{\mu_a, a \in \mathcal{A}\}$ but μ_a is unknown.

$$\mu_a = \tilde{\mu}_{a,t} + \underbrace{(\mu_a - \tilde{\mu}_{a,t})}_{\text{error term}} .$$

Idea

Bound the error term and play a penalized strategy instead.

Towards a better strategy: Simple tools

Lemma (Hoeffding's inequality)

For n i.i.d. random variables $X_i \in [0,1]$ with mean μ, we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geqslant \sqrt{\frac{\ln(1/\delta)}{2n}}\right) \leqslant \delta$$

$$\mathbb{P}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geqslant \sqrt{\frac{\ln(1/\delta)}{2n}}\right) \leqslant \delta\,.$$

UCB strategy

The Upper Confidence Bound algorithm (Auer et al. 2002)

Choose $A_{t+1} = \mathrm{Argmax}\{\mu^+_{a,t}, a \in \mathcal{A}\}$ where

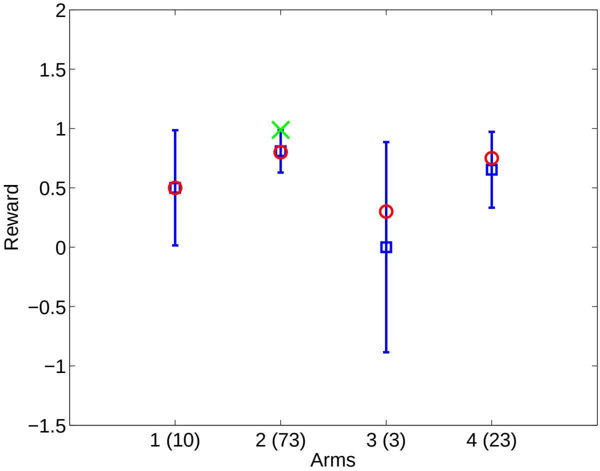
$$\mu^+_{a,t} = \tilde{\mu}_{a,t} + \sqrt{\frac{\ln(1/\delta_t)}{2N_a(t)}} \quad \text{with } \tilde{\mu}_{a,t} = \frac{1}{N_a(t)} \sum_{i=1}^{N_a(t)} X_{i,a}\,.$$

- Choice $\delta_t = t^{-2}(t+1)^{-1}$ gives for each $a \in \mathcal{A}$, $t > A$,

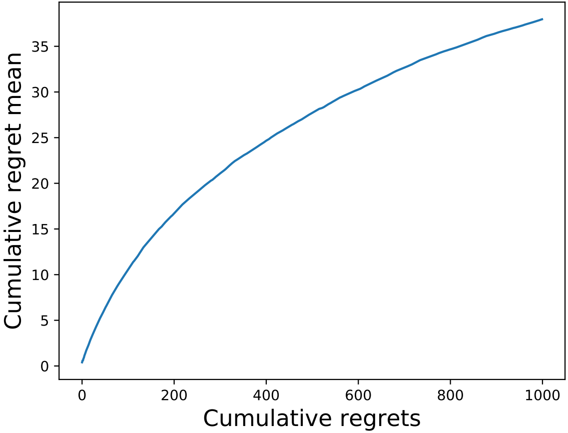
$$\mathbb{P}\left(\mu_a - \tilde{\mu}_{a,t} \geqslant \sqrt{\frac{\ln(1/\delta_t)}{2N_a(t)}}\right) \leqslant \frac{1}{t(t+1)}\,.$$

- "Optimistic strategy"

The Upper–Confidence Bound (UCB) Algorithm

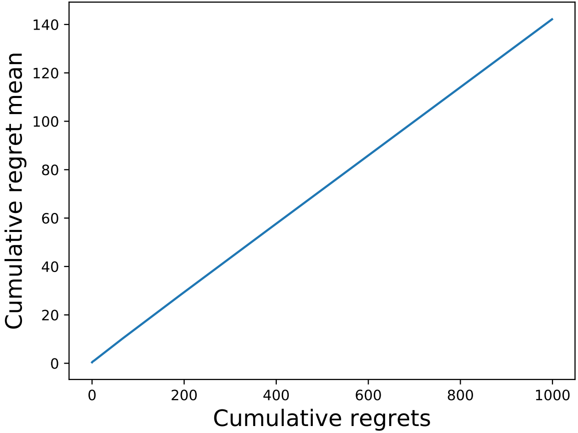


Regret of UCB for a $[\mathcal{B}(0.2), \mathcal{B}(0.4), \mathcal{B}(0.6)]$-bandit



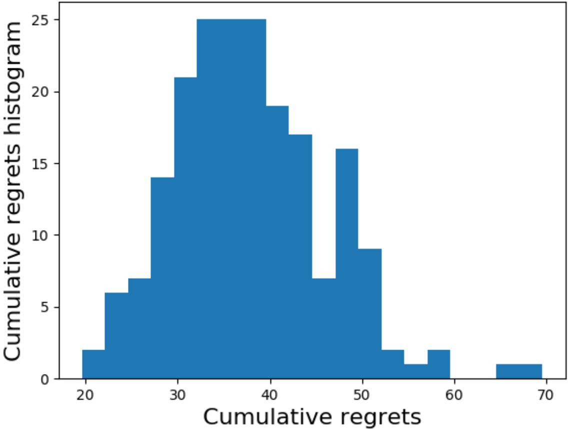
Results averaged over 200 runs.

Regret of FTL for a $[\mathcal{B}(0.2), \mathcal{B}(0.4), \mathcal{B}(0.6)]$-bandit

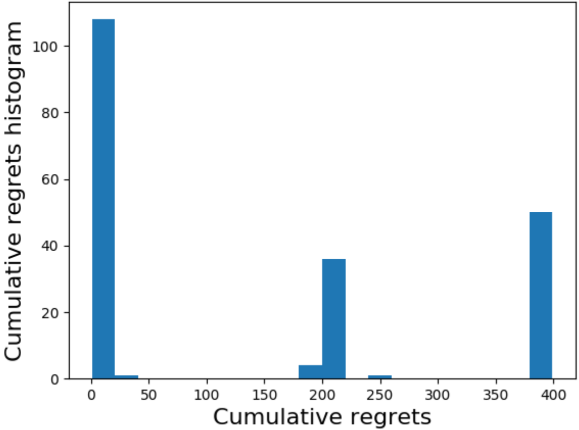


Results averaged over 200 runs.

Regret of UCB for a $[\mathcal{B}(0.2), \mathcal{B}(0.4), \mathcal{B}(0.6)]$-bandit



Regret of FTL for a $[\mathcal{B}(0.2), \mathcal{B}(0.4), \mathcal{B}(0.6)]$-bandit



The Exploration-Exploitation dilemma

$$\mu^+_{a,t} = \tilde{\mu}_{a,t} + \sqrt{\frac{\ln(1/\delta_t)}{2N_a(t)}}\,.$$

Exploitation: "Follow current knowledge"

Choose arm with highest empirical mean: $\tilde{\mu}_{a,t}$

Exploration: Maximally improve current knowledge

Choose least known arm: arm with smallest $N_a(t)$.

The Upper Confidence Bound (UCB) strategy

Assume rewards generated by ν are bounded in $[0, 1]$.

Theorem (Distribution-dependent regret bounds for UCB)

In the stochastic multi-armed bandit game, the UCB strategy with $\delta_t = t^{-2}(t+1)^{-1}$ satisfies the following performance bound.

$$\mathfrak{R}_\nu(T, UCB) \leqslant \sum_{a;\Delta_a>0} \left[\frac{6}{\Delta_a}\ln(T) + 3\Delta_a\right]$$

Scaling in $\displaystyle\sum_{a;\Delta_a>0} \frac{\ln(T)}{\Delta_a}$

Introduction

Multi-armed bandits

Regret lower-bounds

Near-optimal strategies

Boundary crossing for regret analysis

Lower performance bounds

Definition (Uniformly good strategy)

A strategy is uniformly good on $\mathcal{D}$ if for any stochastic bandit $\nu = (\nu_a)_{a\in\mathcal{A}} \in \mathcal{D}$,

$$a \notin \mathcal{A}_\star(\nu) \implies \forall \alpha \in (0,1) \quad \mathbb{E}_\nu[N_a(T)] = o(T^\alpha)\,.$$

Theorem (Lai & Robbins, 1985)

Any uniformly good strategy on the set of *Bernoulli* bandit $\nu = (\mathcal{B}(\theta_1), \dots, \mathcal{B}(\theta_A))$ with means $\theta_a < 1$ must satisfy:

$$a \notin \mathcal{A}_\star(\nu) \implies \liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\ln(T)} \geqslant \frac{1}{\mathtt{KL}(\theta_a, \theta_\star)}\,.$$

$$\text{Thus} \quad \liminf_{T\to\infty} \frac{\mathcal{R}_T(\theta, \pi)}{\ln(T)} \geqslant \sum_{a:\Delta_a>0} \frac{\mu_\star - \mu_a}{\mathtt{KL}(\theta_a, \theta_\star)}\,.$$

Change of measure

- Let $\mathbf{a} = (a_{t'})_{t' \leqslant t}$ be a deterministic sequence of actions.
- For $\nu = (\nu_a)_{a \in \mathcal{A}}$, form $\nu_{\mathbf{a}} = \otimes_{t'=1}^{t} \nu_{a_{t'}}$ on $\mathcal{X}^t$.
- Consider the random variable $Y = (Y_{t'})_{t' \leqslant n}$ in $\mathcal{X}^t$.

$$\ln\left(\frac{d\tilde{\nu}_{\mathbf{a}}}{d\nu_{\mathbf{a}}}(Y)\right) = \sum_{a' \in \mathcal{A}} \sum_{t'=1}^{t} \ln\left(\frac{d\tilde{\nu}_{a'}}{d\nu_{a'}}(Y_{t'})\right) \mathbb{I}\{a_{t'} = a'\}\,.$$

In particular,

- $\forall a' \in \mathcal{A} \backslash \{a\}, \tilde{\nu}_{a'} = \nu_{a'} \implies \ln\left(\frac{d\tilde{\nu}_{\mathbf{a}}}{d\nu_{\mathbf{a}}}(Y)\right) = \sum_{i=1}^{N_a(t)} \ln\left(\frac{d\tilde{\nu}_a}{d\nu_a}(X_{a,i})\right)$
- $\mathbb{E}_{\tilde{\nu}}\left[\ln\left(\frac{d\tilde{\nu}_{\mathbf{a}}}{d\nu_{\mathbf{a}}}(Y)\right)\right] = \sum_{a \in \mathcal{A}} N_a(t) \mathtt{KL}(\tilde{\nu}_a, \nu_a)$

Sketch of proof

- Most confusing environment:
  For $a \notin \mathcal{A}_\star(\nu)$, find $\tilde{\nu}$ such that $a = \mathcal{A}_\star(\tilde{\nu})$.
- Change of measure / Likelihood ratio.
- Asymptotic Maximal Hoeffding inequality.

1. Reduction

$$\frac{\mathbb{E}[N_a(T)]}{\ln(T)} \geqslant c\mathbb{P}_\nu(N_a(T) \geqslant c\ln(T)) \quad \text{(Markov inequality)}$$

Study $\Omega = \{N_T(a) < c\ln(T)\}$. Show that $\mathbb{P}_\nu(\Omega) \to 0$ with T.

2. Confusing instance

Let $\tilde{\nu} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_A)$ be a **maximally confusing** instance for $a \notin \mathcal{A}^\star(\nu)$

$$\begin{cases} \tilde{\theta}_{a'} = \theta_{a'} & \text{if } a' \neq a \\ \tilde{\theta}_a = \lambda & \text{where } \lambda > \mu_\star \text{ (hence } a \in \mathcal{A}_\star(\tilde{\nu})) \end{cases}$$

3. (Bernoulli) log-Likelihood threshold

$$\text{Let } \mathcal{E} = \{\mathcal{L}_{N_a(T)} \leqslant (1-\alpha)\ln(T)\}$$

$$\text{where } \mathcal{L}_m = \sum_{j=1}^{m} \ln\left(\frac{d\nu_{\theta_a}}{d\nu_\lambda}(X_{a,j})\right) \text{ with } d\nu_\theta(x) = \theta^x(1-\theta)^{1-x}\,.$$

$$\begin{aligned}\mathbb{P}_\nu(\Omega\cap\mathcal{E}) &= \mathbb{E}_\nu\left(e^{\ln\left(\frac{d\nu}{d\tilde{\nu}}(Y)\right)}\mathbb{I}\{\Omega\cap E\}\right)\\ &\leqslant T^{1-\alpha}\mathbb{P}_{\tilde{\nu}}(\Omega\cap\mathcal{E}) \quad \text{(Change of measure)}\end{aligned}$$

$$\begin{aligned}\mathbb{P}_\nu(\Omega\cap\mathcal{E}) &\leqslant T^{1-\alpha}\mathbb{P}_{\tilde{\nu}}\Big(\sum_{a'\neq a} N_{a'}(T) > T - c\ln(T)\Big) \quad \Big(\sum_{a'} N_{a'}(T) = T\Big)\\ &\leqslant T^{1-\alpha}\frac{\sum_{a'\neq a}\mathbb{E}_{\tilde{\nu}}[N_{a'}(T)]}{T - c\ln(T)} \quad \text{(Markov inequality)}\\ &= o(1) \quad (\text{Consistency for } \tilde{\nu})\end{aligned}$$

4. (Maximal) concentration inequality

$$\mathbb{P}_\nu(\Omega \cap \mathcal{E}^c) \leqslant \mathbb{P}_\nu\left(\exists m < c\ln(T) : \sum_{j=1}^{m} \underbrace{\ln\left(\frac{d\nu_{\theta_a}(X_{a,j})}{d\nu_\lambda(X_{a,j})}\right)}_{Z_j} > (1-\alpha)\ln(T)\right).$$

$$= \mathbb{P}_\nu\left(\frac{\max_{m<c\ln(T)} \sum_{j=1}^{m} Z_j}{c\ln(T)} > \frac{1-\alpha}{c\,\mathtt{kl}(\theta_a,\lambda)} \underbrace{\mathtt{kl}(\theta_a,\lambda)}_{\mathbb{E}_\theta[Z_j]}\right)$$

> Lemma (Asymptotic maximal Hoeffding inequality)
> For any i.i.d. bounded Z_j with **positive** mean μ,
> $$\forall \eta > 0,\ \lim_{n\to\infty} \mathbb{P}_\nu\left(\frac{\max_{m<n} \sum_{j=1}^{m} Z_j}{n} > (1+\eta)\mu\right) = 0.$$

$\implies$ e.g. $c = \dfrac{1-2\alpha}{\mathtt{kl}(\theta_a,\lambda)}$ to conclude.

Alternative proof

We make use of the fundamental lemma for change of measure:

(Kaufmann, PhD), (Garivier et al. 2016), (Wald 1945)

For a (random) sequence generated by a sequential sampling policy,

$$\mathtt{KL}(\nu_{\mathbf{a}},\tilde{\nu}_{\mathbf{a}}) = \sum_{a'\in\mathcal{A}} \mathbb{E}_\nu[N_{a'}(T)]\mathtt{KL}(\nu_{a'},\tilde{\nu}_{a'}) \geqslant \sup_{\Omega} \mathtt{kl}(\mathbb{P}_\nu[\Omega],\mathbb{P}_{\tilde{\nu}}[\Omega])\,.$$

where $\mathtt{kl}(x,y) = \mathtt{KL}(\mathcal{B}(x),\mathcal{B}(y))$.

Hence $\forall a \notin \mathcal{A}^\star(\nu)$

$$\mathbb{E}_\nu[N_a(T)] \geqslant \sup_{\Omega,\tilde{\nu}} \frac{\mathtt{kl}(\mathbb{P}_\nu[\Omega],\mathbb{P}_{\tilde{\nu}}[\Omega]) - \sum_{a'\neq a}\mathtt{KL}(\nu_{a'},\tilde{\nu}_{a'})\mathbb{E}_\theta[N_{a'}(T)]}{\mathtt{KL}(\nu_a,\tilde{\nu}_a)}\,.$$

$$\mathbb{E}_\nu[N_a(T)] \geqslant \sup_{\Omega,\tilde{\nu}} \frac{\mathtt{kl}(\mathbb{P}_\nu[\Omega], \mathbb{P}_{\tilde{\nu}}[\Omega]) - \sum_{a' \neq a} \mathtt{KL}(\nu_{a'}, \tilde{\nu}_{a'}) \mathbb{E}_\theta[N_{a'}(T)]}{\mathtt{KL}(\nu_a, \tilde{\nu}_a)} .$$

$$\mathbb{E}_\nu[N_a(T)] \geqslant \sup_{\Omega,\tilde\nu} \frac{\mathtt{kl}(\mathbb{P}_\nu[\Omega], \mathbb{P}_{\tilde\nu}[\Omega]) - \sum_{a'\neq a} \mathrm{KL}(\nu_{a'}, \tilde\nu_{a'})\mathbb{E}_\theta[N_{a'}(T)]}{\mathrm{KL}(\nu_a, \tilde\nu_a)}.$$

Choose $\tilde\nu$ such that $\mathcal{A}^\star(\tilde\nu) = \{a\}$, $\Omega = \{N_a(T) > T^\alpha\}$:

- $\mathbb{P}_\nu[\Omega] \leqslant \mathbb{E}_\nu[N_a(T)]T^{-\alpha} = o(1)$
- $\mathtt{kl}(\mathbb{P}_\nu[\Omega], \mathbb{P}_{\tilde\nu}[\Omega]) \simeq \ln\left(\frac{1}{\mathbb{P}_{\tilde\nu}(N_T(a) \leqslant T^\alpha)}\right) \geqslant \ln\left(\frac{T - T^\alpha}{\sum_{a'\neq a} \mathbb{E}_{\tilde\nu}[N_T(a')]}\right) \simeq \ln(T)$.
- Choose $\tilde\nu_{a'}$ for $a' \neq a$: $\tilde\nu_{a'} = \nu_{a'}$ (no constraint)

$$\mathbb{E}_\nu[N_a(T)] \geqslant \sup_{\Omega,\tilde{\nu}} \frac{\mathtt{kl}(\mathbb{P}_\nu[\Omega],\mathbb{P}_{\tilde{\nu}}[\Omega]) - \sum_{a'\neq a}\mathrm{KL}(\nu_{a'},\tilde{\nu}_{a'})\mathbb{E}_\theta[N_{a'}(T)]}{\mathrm{KL}(\nu_a,\tilde{\nu}_a)}\,.$$

Choose $\tilde{\nu}$ such that $\mathcal{A}^\star(\tilde{\nu}) = \{a\}$, $\Omega = \{N_a(T) > T^\alpha\}$:

- $\mathbb{P}_\nu[\Omega] \leqslant \mathbb{E}_\nu[N_a(T)]T^{-\alpha} = o(1)$
- $\mathtt{kl}(\mathbb{P}_\nu[\Omega],\mathbb{P}_{\tilde{\nu}}[\Omega]) \simeq \ln\left(\frac{1}{\mathbb{P}_{\tilde{\nu}}(N_T(a)\leqslant T^\alpha)}\right) \geqslant \ln\left(\frac{T-T^\alpha}{\sum_{a'\neq a}\mathbb{E}_{\tilde{\nu}}[N_T(a')]}\right) \simeq \ln(T)$.
- Choose $\tilde{\nu}_{a'}$ for $a' \neq a$: $\tilde{\nu}_{a'} = \nu_{a'}$ (no constraint)

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\nu[N_a(T)]}{\ln(T)} \geqslant \frac{1-0}{\inf_{\tilde{\nu}_a}\{\mathrm{KL}(\nu_a,\tilde{\nu}_a) : \tilde{\mu}_a > \mu_\star(\nu)\}}$$

Regret lower bounds

This generalizes beyond Bernoulli distributions:

Lower bound (Burnetas & Katehakis, 96)

Any uniformly good strategy on a product set $\mathcal{D} \in \otimes_{a\in\mathcal{A}}\mathcal{D}_a$ of distributions (under mild assumptions) must satisfy

$$\liminf_{T\to\infty} \frac{\mathfrak{R}_T}{\ln T} \geqslant \sum_{a\in\mathcal{A}} \frac{\Delta_a}{\mathcal{K}_a(\nu_a, \mu_\star)}, \quad \mathcal{K}_a(\nu_a, \mu_\star) = \inf_{\nu\in\mathcal{D}_a, \mu_\nu > \mu_\star} \mathtt{KL}(\nu_a, \nu)$$

- Even though the initial problem involves means only, the lower bound depend on the full distributions.

Introduction

Multi-armed bandits

Regret lower-bounds

Near-optimal strategies

Boundary crossing for regret analysis

Historical notes on stochastic bandits and KL-ucb

- 1933 Thompson: Clinical trials. Thompson (1935), Wald (1945).
- 1952 Robbins: Formulation of MABs.
- 1979 Gittins : Optimal strategies as **dynamic allocation indices**.
- 1985 Lai&Robbins: Indices as **upper confidence bounds**. Asymptotically optimal policies see also Burnetas&Katehakis (1997), Agrawal (1995).
- 1987 Lai: The KL-ucb algorithm.
- 2002 Auer, Cesa-Bianchi, Fischer: First **finite-time** regret analysis.
- 2010 Honda&Takemura: Novel view on asymptotically optimal strategies.
- 2011 M., Munos, Stoltz: KL-ucb **finite-time** analysis for discrete distributions; Cappe&Garivier (2011): Bernoulli distributions.
- 2013 Cappe, Garivier, M. Munos, Stoltz: KL-ucb for **dimension 1 exponential families** and discrete distributions.

Historical notes on stochastic bandits and KL-ucb

- 1933 Thompson: Clinical trials. Thompson (1935), Wald (1945).
- 1952 Robbins: Formulation of MABs.
- 1979 Gittins : Optimal strategies as **dynamic allocation indices**.
- 1985 Lai&Robbins: Indices as **upper confidence bounds**. Asymptotically optimal policies see also Burnetas&Katehakis (1997), Agrawal (1995).
- **1988** Lai: Boundary crossing probabilities for exponential families.
- 2002 Auer, Cesa-Bianchi, Fischer: First **finite-time** regret analysis.
- 2010 Honda&Takemura: Novel view on asymptotically optimal strategies.
- 2011 M., Munos, Stoltz: KL-ucb **finite-time** analysis for discrete distributions; Cappe&Garivier (2011): Bernoulli distributions.
- 2013 Cappe, Garivier, M. Munos, Stoltz: KL-ucb for **dimension 1 exponential families** and discrete distributions.

A strategy inspired from lower bounds

- Lower bound not only provides **limiting regret** performance. It shows that in order to be **uniformly optimal** on a set of bandit configurations $\mathcal{D}$, sub-optimal arms have to be pulled some amount of time:

$$\mathbb{E}[N_a(T)]\mathtt{KL}(\nu_a, \tilde{\nu}_a) \geqslant \ln(T) \text{ as } T \to \infty, \text{ when } a \in A_\star(\tilde{\nu})$$

- KL-UCB: Consider $\{\tilde{\nu}_a : N_a(t)\mathtt{KL}(\nu_a, \tilde{\nu}_a) \leqslant \ln(t)\}$
- Pulling a increases $N_a(t)$ by one, thus possibly reduces this set: try to **remove** the environment with largest mean reward.

The class of KL-ucb algorithms

Use empirical distributions: $\widehat{\nu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} \delta_{Y_s} \mathbb{I}_{\{a_s = a\}}$.

KL-ucb for a family $\mathcal{D}$ (generic form)

Pick arm $\quad a_{t+1} \in \underset{a \in \mathcal{A}}{\mathrm{Argmax}}\ U_a(t)$ where

$$U_a(t) = \sup\Big\{\tilde{\mu}_a : \ \tilde{\nu} \in \mathcal{D}_a \text{ and } N_a(t)\mathtt{KL}\Big(\Pi_{\mathcal{D}}(\widehat{\nu}_a(t)), \tilde{\nu}\Big) \leqslant f(t)\Big\}.$$

with Operator $\Pi_{\mathcal{D}} : \mathcal{P}(\mathbb{R}) \to \mathcal{D}$; Non-decreasing $f : \mathbb{N} \to \mathbb{R}$

Rewriting lemma (Cappe et al., 2013)

Under mild assumption on $\mathcal{D} \subset \mathcal{P}([\mu^-, \mu^+])$,

$$U_a(t) = \max\left\{\tilde{\mu} \in [\mu^-, \mu^+) : \ \mathcal{K}_a\Big(\Pi_{\mathcal{D}}\left(\widehat{\nu}_a(t)\right), \tilde{\mu}\Big) \leqslant \frac{f(t)}{N_a(t)}\right\}.$$

The class of KL-ucb algorithms

Use empirical distributions: $\widehat{\nu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} \delta_{Y_s} \mathbb{I}_{\{a_s=a\}}$.

KL-ucb+ for a family $\mathcal{D}$ (generic form)

Pick arm $a_{t+1} \in \operatorname{Argmax}_{a\in\mathcal{A}} U_a(t)$ where

$$U_a(t)=\sup\Big\{\tilde{\mu}_a : \tilde{\nu}\in\mathcal{D}_a \text{ and } N_a(t)\mathrm{KL}\Big(\Pi_{\mathcal{D}}(\widehat{\nu}_a(t)),\tilde{\nu}\Big)\leqslant f\Big(\frac{t}{N_a(t)}\Big)\Big\}.$$

with Operator $\Pi_{\mathcal{D}} : \mathcal{P}(\mathbb{R}) \to \mathcal{D}$; Non-decreasing $f : \mathbb{N} \to \mathbb{R}$

Rewriting lemma (Cappe et al., 2013)

Under mild assumption on $\mathcal{D} \subset \mathcal{P}([\mu^-, \mu^+])$,

$$U_a(t)=\max\Big\{\tilde{\mu} \in [\mu^-, \mu^+) : \mathcal{K}_a\Big(\Pi_{\mathcal{D}}(\widehat{\nu}_a(t)), \tilde{\mu}\Big) \leqslant \frac{f(t/N_a(t))}{N_a(t)}\Big\}.$$

KL-UCB: Class of distributions

The strategy depends on the considered class $\mathcal{D}$. Example of $\mathcal{D}$:

- Bernoulli: $\nu_\theta = \mathcal{B}(\theta)$
- Standard Gaussian: $\nu_\theta = \mathcal{N}(\theta, 1)$
- Exponential family of dimension 1:

$$\{\nu_\theta \in \mathcal{P}(\mathcal{X}) : \forall x \in \mathcal{X} \ \nu_\theta(x) = \exp\left(\theta x - \psi(\theta)\right)\nu_0(x),\ \theta \in \mathbb{R}\},$$

Exponential families of higher dimension

The exponential family $\mathcal{E}(F;\nu_0)$ generated by the function F and the reference measure ν_0 on the set $\mathcal{X}$ is

$$\Big\{\nu_\theta \in \mathcal{P}(\mathcal{X}) :\ \forall x \in \mathcal{X}\ \nu_\theta(x) = \exp\big(\langle \theta, F(x)\rangle - \psi(\theta)\big)\nu_0(x),\ \theta \in \mathbb{R}^K\Big\},$$

with

Exponential families of higher dimension

The exponential family $\mathcal{E}(F;\nu_0)$ generated by the function F and the reference measure ν_0 on the set $\mathcal{X}$ is

$$\Big\{\nu_\theta\in\mathcal{P}(\mathcal{X}):\ \forall x\in\mathcal{X}\ \nu_\theta(x)=\exp\big(\langle\theta,F(x)\rangle-\psi(\theta)\big)\nu_0(x),\ \theta\in\mathbb{R}^K\Big\},$$

with

- Log-partition function: $\psi(\theta)\stackrel{\text{def}}{=}\ln\int_{\mathcal{X}}\exp\Big(\langle\theta,F(x)\rangle\Big)\nu_0(dx)$
- Canonical parameter set: $\Theta_{\mathcal{D}}\stackrel{\text{def}}{=}\Big\{\theta\in\mathbb{R}^K:\psi(\theta)<\infty\Big\}$
- Invertible parameter set:
  $\Theta_I\stackrel{\text{def}}{=}\Big\{\theta\in\Theta_{\mathcal{D}}:0<\lambda_{\texttt{MIN}}(\nabla^2\psi(\theta))\leqslant\lambda_{\texttt{MAX}}(\nabla^2\psi(\theta))<\infty\Big\}$
  where $\lambda_{\texttt{MIN}}(M)$ and $\lambda_{\texttt{MAX}}(M)$ are the minimum and maximum eigenvalues of a semi-definite positive matrix M.

Examples

Bernoulli ($K=1,\ F(x)=x$), Gaussian ($K=2,\ F(x)=(x,x^2)$).

Introduction

Multi-armed bandits

Regret lower-bounds

Near-optimal strategies

Boundary crossing for regret analysis

From regret to boundary crossing probabilities

The number of pulls of a sub-optimal arm $a \in \mathcal{A}$ by Algorithm KL-ucb satisfies

$$\mathbb{E}[N_a(T)] \leqslant 2 + \inf_{n_0 \leqslant T} \Big\{ n_0 + \sum_{n \geqslant n_0+1}^{T} \underbrace{\mathbb{P}\Big\{ \widehat{\nu}_{a,n} \in \mathcal{C}_{\mu^\star-\varepsilon}\Big(f(T)/n\Big) \Big\}}_{\text{Finite-time Sanov term}} \Big\}$$

$$+ \sum_{t=|\mathcal{A}|}^{T-1} \underbrace{\mathbb{P}\Big\{ N_{a^\star}(t)\, \mathcal{K}_{a^\star}\big(\Pi_{\mathcal{D}}(\widehat{\nu}_{a^\star, N_{a^\star}(t)}),\, \mu^\star - \varepsilon\big) > f(t) \Big\}}_{\text{Boundary Crossing Probability}} .$$

for any $\varepsilon \in \mathbb{R}^+$ such that $\varepsilon \in (0, \min_{a \in \mathcal{A}\setminus\{a^\star\}} \Delta_a)$, and introducing the (open, convex) set

$$\mathcal{C}_\mu(\gamma) \quad = \quad \Big\{ \nu' \in \mathcal{P}(\mathbb{R}) : \quad \mathcal{K}_a(\Pi_a(\nu'), \mu) < \gamma \Big\} .$$

From regret to boundary crossing probabilities

The number of pulls of a sub-optimal arm $a \in \mathcal{A}$ by Algorithm KL-ucb+ satisfies

$$\mathbb{E}[N_a(T)] \leqslant 2 + \inf_{n_0 \leqslant T} \left\{ n_0 + \sum_{n \geqslant n_0+1}^{T} \underbrace{\mathbb{P}\Big\{ \widehat{\nu}_{a,n} \in \mathcal{C}_{\mu^\star - \varepsilon}\Big(f(T/n)/n \Big) \Big\}}_{\text{Finite-time Sanov term}} \right\}$$

$$+ \sum_{t=|\mathcal{A}|}^{T-1} \underbrace{\mathbb{P}\Big\{ N_{a^\star}(t)\, \mathcal{K}_{a^\star}\big(\Pi_{\mathcal{D}}(\widehat{\nu}_{a^\star, N_{a^\star}(t)}),\, \mu^\star - \varepsilon\big) > f(t/N_{a^\star}(t)) \Big\}}_{\text{Boundary Crossing Probability}} .$$

for any $\varepsilon \in \mathbb{R}^+$ such that $\varepsilon \in (0, \min_{a \in \mathcal{A} \setminus \{a^\star\}} \Delta_a)$, and introducing the (open, convex) set

$$\mathcal{C}_\mu(\gamma) = \Big\{ \nu' \in \mathcal{P}(\mathbb{R}) : \ \mathcal{K}_a(\Pi_a(\nu'), \mu) < \gamma \Big\} .$$

From regret to boundary crossing probabilities: Goal

$$\sum_{t=|\mathcal{A}|}^{T-1} \underbrace{\mathbb{P}\Big\{ N_{a^\star}(t)\, \mathcal{K}_{a^\star}\big(\Pi_{\mathcal{D}}(\widehat{\nu}_{a^\star, N_{a^\star}(t)}),\, \mu^\star - \varepsilon\big) > f\big(t/N_{a^\star}(t)\big) \Big\}}_{\text{Goal: } o(1/t)} = o(\ln(T))$$

From regret to boundary crossing probabilities: Goal

$$\mathbb{P}\Big\{\bigcup_{n=1}^{t} n\,\mathcal{K}_{a^\star}\big(\Pi_{\mathcal{D}}(\widehat{\nu}_{a^\star,n}),\, \mu^\star-\varepsilon\big) > f(t/n)\Big\} = o(1/t)$$

From regret to boundary crossing probabilities: Goal

$$\mathbb{P}_\nu\Big\{\bigcup_{n=1}^{t} n\,\mathcal{K}(\Pi_{\mathcal{D}}(\widehat{\nu}_n),\, E[\nu]-\varepsilon) > f(t/n)\Big\} = o(1/t)$$

Boundary-crossing probabilities

A tribute to T.L. Lai

Boundary crossing probabilities

K-dimensional exponential families

Existing results

Main results

Exponential families

Exponential family

The exponential family $\mathcal{E}(F;\nu_0)$ generated by the function F and the reference measure ν_0 on the set $\mathcal{X}$ is

$$\Big\{\nu_\theta \in \mathfrak{M}_1(\mathcal{X}) : \forall x \in \mathcal{X}\ \nu_\theta(x) = \exp\big(\langle\theta, F(x)\rangle - \psi(\theta)\big)\nu_0(x),\ \theta \in \mathbb{R}^K\Big\},$$

with

- Log-partition function: $\psi(\theta) \stackrel{\text{def}}{=} \ln \int_{\mathcal{X}} \exp\Big(\langle\theta, F(x)\rangle\Big)\nu_0(dx)$
- Canonical parameter set: $\Theta_{\mathcal{D}} \stackrel{\text{def}}{=} \Big\{\theta \in \mathbb{R}^K : \psi(\theta) < \infty\Big\}$
- Invertible parameter set:
  $\Theta_I \stackrel{\text{def}}{=} \Big\{\theta \in \Theta_{\mathcal{D}} : 0 < \lambda_{\texttt{MIN}}(\nabla^2\psi(\theta)) \leqslant \lambda_{\texttt{MAX}}(\nabla^2\psi(\theta)) < \infty\Big\}$
  where $\lambda_{\texttt{MIN}}(M)$ and $\lambda_{\texttt{MAX}}(M)$ are the minimum and maximum eigenvalues of a semi-definite positive matrix M.

Examples

Bernoulli ($K = 1$, $F(x) = x$), Gaussian ($K = 2$, $F(x) = (x, x^2)$).

Useful properties

Bregman divergence

$$\mathtt{KL}(\nu_\theta, \nu_{\theta'}) = \mathcal{B}^\psi(\theta, \theta') \stackrel{\text{def}}{=} \psi(\theta') - \psi(\theta) - \langle \theta' - \theta, \nabla\psi(\theta)\rangle \,.$$

Bregman smoothness property

$$\|\theta - \theta'\| \frac{v_\Theta}{2} \leqslant \mathcal{B}^\psi(\theta, \theta') \leqslant \|\theta - \theta'\| \frac{V_\Theta}{2}$$

where $v_\Theta = \inf_{\theta\in\Theta} \lambda_{\mathtt{MAX}}(\nabla^2\psi(\theta))$, $V_\Theta = \sup_{\theta\in\Theta} \lambda_{\mathtt{MAX}}(\nabla^2\psi(\theta))$.

We can rewrite: $\mathcal{K}(\nu_\theta, \mu) = \inf\{\mathtt{KL}(\nu_\theta, \nu_{\theta'}) : E[\nu_{\theta'}] > \mu\}$.

Boundary crossing probabilities

K-dimensional exponential families

Existing results

Main results

What was known

- Optimality of KL-UCB strategy is only known for specific classes of distributions:
  Bernoulli, Gaussian, exponential families fo dimension 1, Discrete distributions.
- Goal: Exponential families of arbitrary dimension $K > 1$.

Technicalities: large enough sets.

Estimation

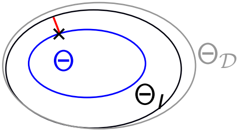
$\widehat{F}_n = \frac{1}{n}\sum_{i=1}^n F(X_i) \in \mathbb{R}^K$, then "$\widehat{\theta}_n = \nabla\psi^{-1}(\widehat{F}_n)$".
(Assumption required, essentially regular family and $\theta^\star \in \mathring{\Theta}_I$)

Enlarged parameter set

The enlargement of size $\rho \in \mathbb{R}^+$ of Θ is defined by

$$\Theta_\rho \stackrel{\text{def}}{=} \left\{\theta \in \mathbb{R}^K ; \inf_{\theta' \in \Theta} \|\theta - \theta'\| < \rho \right\}.$$

Further, let $v_\rho \stackrel{\text{def}}{=} \inf_{\theta \in \Theta_\rho} \lambda_{\texttt{MIN}}(\nabla^2\psi(\theta))$, $V_\rho \stackrel{\text{def}}{=} \sup_{\theta \in \Theta_\rho} \lambda_{\texttt{MAX}}(\nabla^2\psi(\theta))$.



Technicalities: large enough sets.

Estimation

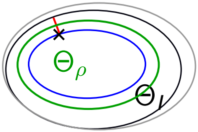
$\widehat{F}_n = \frac{1}{n}\sum_{i=1}^n F(X_i) \in \mathbb{R}^K$, then "$\widehat{\theta}_n = \nabla\psi^{-1}(\widehat{F}_n)$".
(Assumption required, essentially regular family and $\theta^\star \in \mathring{\Theta}_I$)

Enlarged parameter set

The enlargement of size $\rho \in \mathbb{R}^+$ of Θ is defined by

$$\Theta_\rho \stackrel{\text{def}}{=} \left\{\theta \in \mathbb{R}^K \,;\, \inf_{\theta' \in \Theta} \|\theta - \theta'\| < \rho \right\}.$$

Further, let $v_\rho \stackrel{\text{def}}{=} \inf_{\theta\in\Theta_\rho} \lambda_{\texttt{MIN}}(\nabla^2\psi(\theta)),\ V_\rho \stackrel{\text{def}}{=} \sup_{\theta\in\Theta_\rho} \lambda_{\texttt{MAX}}(\nabla^2\psi(\theta))$.



For $\rho = -/2$, when $\widehat{F}_n \in \nabla\psi(\Theta_\rho)$,

$$\exists \widehat{\theta}_n \in \Theta_\rho \subset \mathring{\Theta}_I, \nabla\psi(\widehat{\theta}_n) = \widehat{F}_n.$$

Existing results

Theorem [Cappe et al. 2013]

For the canonical $(F(x)=x)$ exponential family of dimension 1,

$$\mathbb{P}_{\theta^\star}\Big\{\bigcup_{n=1}^{t} n\mathcal{K}(\Pi(\widehat{\nu}_n),\, \mu^\star) > f(t) \cap \mu^\star > \widehat{\mu}_n\Big\} \leqslant e\lceil f(t)\ln(t)\rceil e^{-f(t)}\,.$$

Use $f(x) = \ln(x) + 3\ln\ln(x)$ makes the bound $o(1/t)$.

Existing results

Theorem [Cappe et al. 2013]

For the canonical ($F(x)=x$) exponential family of dimension 1,

$$\mathbb{P}_{\theta^\star}\Big\{\bigcup_{n=1}^{t} n\mathcal{K}(\Pi(\widehat{\nu}_n),\,\mu^\star) > f(t) \cap \mu^\star > \widehat{\mu}_n\Big\} \leqslant e\lceil f(t)\ln(t)\rceil e^{-f(t)}\,.$$

Use $f(x) = \ln(x) + 3\ln\ln(x)$ makes the bound $o(1/t)$.

Theorem [Lai, 1988] (exp. family of dimension K)

Define the cone $\mathcal{C}_p(\theta) = \{\theta' \in \mathbb{R}^K : \langle\theta',\theta\rangle \geqslant p\|\theta\|\|\theta'\|\}$, for $p > 0$.
Let $f(x) = \alpha\ln(x) + \xi\ln\ln(x)$. Then for all $\theta\in\Theta$ such that $|\theta-\theta^\star|^2 \geqslant \delta_t$, where $\delta_t \stackrel{t\to\infty}{\to} 0$, $t\delta_t \stackrel{t\to\infty}{\to} \infty$,

$$\mathbb{P}_{\theta^\star}\Big\{\bigcup_{n=1}^{t}\widehat{\theta}_n\in\Theta_\rho \cap n\mathcal{B}^{\psi}(\widehat{\theta}_n,\theta)\geqslant f\Big(\frac{t}{n}\Big)\cap\overbrace{\nabla\psi(\widehat{\theta}_n)-\nabla\psi(\theta)\in\mathcal{C}_p(\theta-\theta^\star)}^{\text{Cone condition}}\Big\}$$
$$\stackrel{t\to\infty}{=} O\Big(t^{-\alpha}|\theta-\theta^\star|^{-2\alpha}\ln^{-\xi-\alpha+K/2}(t|\theta-\theta^\star|^2)\Big)$$

Discussion

Comparison

[Cappe et al. 2013]	[Lai 1988]
• $f(t)$ (`KL-ucb`)	• $f(t/n)$ (`KL-ucb+`)
• Dimension 1 or discrete	• Dimension K.
• Finite time	• Asymptotic + Cone condition
• $o(1/t)$ requires $\xi > 2$ and $\xi \geqslant 3$	• $o(1/t)$ requires $\xi > K/2 - 1$.

Discussion

Comparison

[Cappe et al. 2013]	[Lai 1988]
• $f(t)$ (`KL-ucb`)	• $f(t/n)$ (`KL-ucb+`)
• Dimension 1 or discrete	• Dimension K.
• Finite time	• Asymptotic + Cone condition
• $o(1/t)$ requires $\xi > 2$ and $\xi \geqslant 3$	• $o(1/t)$ requires $\xi > K/2 - 1$.

[Lai, 1988]: proof based on a **change of measure argument**.
Takes advantage of gap between $\mu^\star$ and $\mu^\star - \varepsilon$.
Proof written for $K = 1$, sketched for general K.

Goals

- remove cone condition: cone covering of the space.
- make it non asymptotic: finite-time concentration.
- fully explicit proof.

A note about cone condition

- Already present in dimension 1:
$$\mathbb{P}_{\theta^\star}\Big\{\bigcup_{n=1}^{t} n\mathcal{K}(\Pi(\widehat{\nu}_n),\, \mu^\star) > f(t) \cap \underbrace{\mu^\star > \widehat{\mu}_n}_{\text{Cone condition !}}\Big\}$$
- Cones are natural objects to define **partial orders** on any structure.
$\mathcal{C}_p(\theta) = \{\theta' \in \mathbb{R}^K : \langle \theta', \theta \rangle \geqslant p\|\theta\|\|\theta'\|\}$ is a (convex, pointed, salient) cone and induces such a partial order on $\mathbb{R}^K$.
- Cones are one of the most powerful geometric objects in maths.

Main result overview (informal statement)

Theorem (Informal)

Let $f(x) = \ln(x) + \xi \ln\ln(x)$. *Let* $\mathcal{D}$ *be an exponential family:*

$$\Big\{\nu_\theta : \forall x \in \mathcal{X}\ \nu_\theta(x) = \exp\big(\langle\theta, F(x)\rangle - \psi(\theta)\big)\nu_0(x),\ \theta \in \mathbb{R}^K\Big\},$$

with parameter function $F : \mathcal{X} \to \mathbb{R}^K$ *and reference measure* ν_0. *Then, under some mild condition on* $\mathcal{D}$, *it holds* $\forall \varepsilon \in \mathbb{R}^+, \forall t \in \mathbb{N}$

$$\mathbb{P}\Big\{\bigcup_{n=1}^{t} n\,\mathcal{K}(\Pi_{\mathcal{D}}(\widehat{\nu}_n), E[\nu] - \varepsilon) > f(t)\Big\} \leqslant \frac{C}{t}\ln(t)^{K/2-\xi} e^{-c\sqrt{f(t)}},$$

with c, C *explicit (small) constants depending on* $\mathcal{D}$ *and* ε.

We recommend in practice: $\xi \simeq (K/2 - 2c)_+$ or $(K-1)/2$.

Main result overview (informal statement)

Theorem (Informal)

Let $f(x) = \ln(x) + \xi \ln\ln(x)$. *Let* $\mathcal{D}$ *be an exponential family:*

$$\left\{\nu_\theta :\ \forall x \in \mathcal{X}\ \nu_\theta(x) = \exp\big(\langle\theta, F(x)\rangle - \psi(\theta)\big)\nu_0(x),\ \theta \in \mathbb{R}^K\right\},$$

with parameter function $F : \mathcal{X} \to \mathbb{R}^K$ *and reference measure* ν_0. *Then, under some mild condition on* $\mathcal{D}$, *it holds* $\forall \varepsilon \in \mathbb{R}^+, \forall t \geqslant t_0$

$$\mathbb{P}\left\{\bigcup_{n=1}^{t} n\,\mathcal{K}(\Pi(\widehat{\nu}_n), E[\nu] - \varepsilon) > f(t/n)\right\} \leqslant \frac{C}{t}\ln(tc)^{K/2-\xi-1},$$

with c, C, t_0 *explicit (small) constants depending on* $\mathcal{D}$ *and* ε.

This suggests to tune ξ as: $\xi \simeq (K/2 - 1)_+$.

Boundary crossing probabilities

K-dimensional exponential families

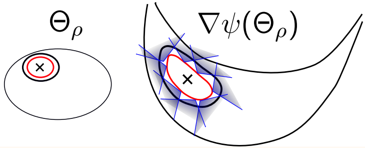
Existing results

Main result

Main results I

- For $\varepsilon > 0$, let $\rho_\varepsilon = \inf\{\|\theta' - \theta\| : \mu_{\theta'} = \mu^\star - \varepsilon, \mu_\theta = \mu^\star\}$.
- Let $C_{p,\eta,K}$ be the cone-covering number of $\nabla\psi(\Theta_\rho \backslash \mathcal{B}_2(\theta^\star; \rho_\varepsilon))$

with minimal angular separation p, excluding $\nabla\psi(\Theta_\rho \backslash \mathcal{B}_2(\theta^\star; \eta\rho_\varepsilon))$.



For all $\eta < 1$, $C_{p,\eta,K} = O((1-p)^{-K})$, $C_{p,\eta,K} \stackrel{\eta\to 1}{\to} \infty$; $C_{p,\eta,1} = 2$.

- Let $\chi = p\eta\sqrt{2{v_\rho}^2/V_\rho}$ and

$$C = C_{p,\eta,K}\Big(2\max\Big\{\frac{8{V_\rho}^4}{p\rho^2 v_\rho^6}, \frac{{V_\rho}^3}{{v_\rho}^4}, \frac{16{V_\rho}^5}{p{v_\rho}^6(\frac{1}{2}+\frac{1}{K})}\Big\}^{K/2} + 1\Big).$$

For Bernoulli with means $\mu \in [\mu_\rho, 1-\mu_\rho]$: $C \leqslant \frac{1}{4\mu_\rho^3(1-\mu_\rho)^3} + 2$.

Main results

Main result for $f(t)$

For all $\rho < \rho^\star$ and all t such that $f(t) \geqslant 1$ it holds

$$\mathbb{P}_{\theta^\star}\Big\{ \bigcup_{1 \leqslant n < t} \widehat{\theta}_n \in \Theta_\rho \cap \mathcal{K}(\Pi(\widehat{\nu}_n), \mu^\star - \varepsilon) \geqslant f(t)/n \Big\} \leqslant$$

$$\frac{C(1+\frac{1}{\chi\rho_\varepsilon})}{t}\Big(1+\xi\frac{\ln\ln(t)}{\ln(t)}\Big)^{K/2} \ln(t)^{-\xi+K/2} e^{-\chi\rho_\varepsilon\sqrt{\ln(t)+\xi\ln\ln(t)}}.$$

We recommend $\xi > K/2 - 2\chi\rho_\varepsilon$ since otherwise the asymptotic regime of $\chi\rho_\varepsilon\sqrt{\ln(t)} - (K/2-\xi)\ln\ln(t)$ may take a massive amount of time to kick-in. In practice $\xi = K/2 - 1/2$ is interesting, since $\ln(t)^{K/2-\xi} = \sqrt{\ln(t)} < 5$ for all $t \leqslant 10^9$.

Main result for $f(t/n)$

Main result for $f(t/n)$

For all $\rho < \rho^\star$, it holds for $\xi \geqslant (K/2-1)_+$ and $t \geqslant 85\chi^{-2}$,

$$\mathbb{P}_{\theta^\star}\Big\{\bigcup_{n=1}^{t} \widehat{\theta}_n \in \Theta_\rho \cap \mathcal{K}(\Pi(\widehat{\nu}_n), \mu^\star - \varepsilon) \geqslant f(t/n)/n\Big\} \leqslant C\Bigg[e^{-\chi\rho_\varepsilon\sqrt{tf(4)/4}} +$$

$$\frac{(1+\xi)^{K/2}}{ct\ln(tc)}\begin{cases}\frac{16}{3}\ln\big(tc\frac{\ln(tc)}{4}\big)^{K/2-\xi} + 80\ln(1.25)^{K/2-\xi} & \text{if } \xi \geqslant K/2\\ \frac{16}{3}\ln\big(\frac{t}{3}\big)^{K/2-\xi} + 80\ln\big(t\frac{c\ln(tc)}{4-c\ln(tc)}\big)^{K/2-\xi} & \text{if } \xi \in [\frac{K}{2}-1, \frac{K}{2}]\end{cases}\Bigg],$$

$$o\Big(\ln(t)\Big) \text{ if } \xi > \frac{K}{2} - 1$$

where $c = \frac{\rho_\varepsilon^2\chi^2}{4\ln(5)^2}$.

Practical consequences

The restriction to $t \geqslant 85\chi_\varepsilon^{-2}$ is merely for $\xi \simeq K/2-1$. It is less restrictive as ξ gets larger. For $\xi \geqslant K/2$, it becomes $t \geqslant 76\chi_\varepsilon^{-2}$.

Critical value

$K/2-1$ (when non-negative) is a critical value for ξ: bounds on boundary crossing probabilities are summable in t iff $\xi > K/2-1$.
In practice we recommend ξ to be away from $K/2-1$.

Adequacy with experiments

When $K=1$, $\max(K/2-1, 0)=0$: sharp phase transition observed for `KL-ucb+` precisely at $\xi=0$: Linear regret for $\xi<0$ and logarithmic regret for $\xi=0$.
For `KL-ucb`, smooth transition at $\xi=0$ depending on the problem.

Boundary crossing probabilities

K-dimensional exponential families

Existing and novel results

Proof techniques

Main ideas of the proof

- **Peeling** argument: sandwich $N(t) \in [n_i, n_{i+1})$, $i \in \mathbb{N}$.

new **Cone** covering: to localize $\widehat{\theta}_n$ outside of $\mathcal{B}_2(\theta^\star, \rho_\varepsilon)$; introduce points $(\theta^\star_c)_{c \leqslant C}$ and (dual) cones $\mathcal{C}(\theta^\star_c)$.

- Double change of measure: 1) from $\theta^\star$ to θ_c, then 2) from $\theta^\star_c$ to the ball $\nabla\psi^{-1}(\mathcal{B}_2(\nabla\psi(\theta^\star_c), \eta) \cap \mathcal{C}(\theta^\star_c))$.
- **Bregman** divergence and Hessian: explicit computations.

new **Concentration** and boundary effects: finite-time concentration inside a cone.

◇ Tight handling of peeling ratios: from $\xi \simeq K/2$ to $K/2 - 1$.

Peeling and covering

Let $\beta, \eta \in (0,1), b > 1$ and define $l_t = \lceil \ln_b(\beta(t+1)) \rceil$. Then

$$\mathbb{P}_{\theta^\star}\Big\{ \bigcup_{1 \leqslant n \leqslant t} \widehat{\theta}_n \in \Theta_\rho \cap \mathcal{K}(\Pi(\widehat{\nu}_n), \mu^\star - \varepsilon) \geqslant f(t/n)/n \Big\} \leqslant$$

Peeling and covering

Let $\beta, \eta \in (0,1)$, $b > 1$ and define $I_t = \lceil \ln_b(\beta(t+1)) \rceil$. Then

$$\mathbb{P}_{\theta^\star}\Big\{ \bigcup_{1 \leqslant n \leqslant t} \widehat{\theta}_n \in \Theta_\rho \cap \mathcal{K}(\Pi(\widehat{\nu}_n), \mu^\star - \varepsilon) \geqslant f(t/n)/n \Big\} \leqslant$$

$$\sum_{i=0}^{I_t - 1} \sum_{c=1}^{C} \mathbb{P}_{\theta^\star}\Big\{ \bigcup_{n=b^i}^{b^{i+1}-1} \overbrace{\widehat{\theta}_n \in \Theta_\rho \cap \widehat{F}_n \in \mathcal{C}_p(\theta^\star_c) \cap \mathcal{B}^\psi(\widehat{\theta}_n, \theta^\star_c) \geqslant \frac{f(t/n)}{n}}^{E_{c,p}(n,t)} \Big\},$$

Peeling and covering

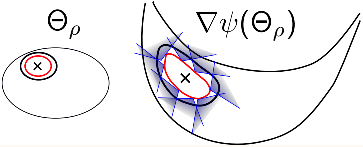
Let $\beta, \eta \in (0,1)$, $b > 1$ and define $I_t = \lceil \ln_b(\beta(t+1)) \rceil$. Then

$$\mathbb{P}_{\theta^\star}\Big\{ \bigcup_{1 \leqslant n \leqslant t} \widehat{\theta}_n \in \Theta_\rho \cap \mathcal{K}(\Pi(\widehat{\nu}_n), \mu^\star - \varepsilon) \geqslant f(t/n)/n \Big\} \leqslant$$

$$\sum_{i=0}^{I_t - 1} \sum_{c=1}^{C} \mathbb{P}_{\theta^\star}\Big\{ \bigcup_{n=b^i}^{b^{i+1}-1} \overbrace{\widehat{\theta}_n \in \Theta_\rho \cap \widehat{F}_n \in \mathcal{C}_p(\theta^\star_c) \cap \mathcal{B}^\psi(\widehat{\theta}_n, \theta^\star_c) \geqslant \frac{f(t/n)}{n}}^{E_{c,p}(n,t)} \Big\},$$

where $C = C_{p,\eta,K}$ cone covering number of $\nabla\psi(\Theta_\rho \setminus \mathcal{B}_2(\theta^\star, \rho_\varepsilon))$ with cones $\forall c \leqslant C, \mathcal{C}_p(\theta^\star_c) := \mathcal{C}_p(\nabla\psi(\theta^\star_c); \theta^\star - \theta^\star_c)$, $\theta^\star_c \notin \mathcal{B}_2(\theta^\star, \eta\rho_\varepsilon)$,

where $\mathcal{C}_p(y; \Delta) = \Big\{ y' \in \mathbb{R}^K : \langle y' - y, \Delta \rangle \geqslant p \|y' - y\| \|\Delta\| \Big\}$:



For all $\eta < 1$, $C_{p,\eta,K} = O((1-p)^{-K})$, $C_{p,\eta,K} \overset{\eta\to 1}{\to} \infty$; $C_{p,\eta,1} = 2$.

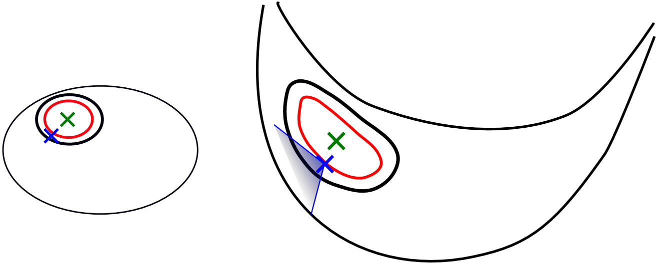
First change of measure

Change of measure

If $n \to nf(t/n)$ is non-decreasing, then for any increasing sequence $\{n_i\}_{i\geqslant 0}$ of non-negative integers it holds

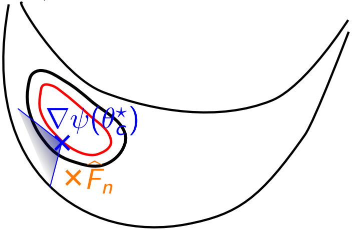
$$\mathbb{P}_{\theta^\star}\Big\{\bigcup_{n=n_i}^{n_{i+1}-1} E_{c,p}(n,t)\Big\} \leqslant \exp\Big(-n_i\alpha^2 - \chi\sqrt{n_i f\big(\frac{t}{n_i}\big)}\Big)\mathbb{P}_{\theta_c^\star}\Big\{\bigcup_{n=n_i}^{n_{i+1}-1} E_{c,p}(n,t)\Big\}$$

where $\alpha = \eta\rho_\varepsilon\sqrt{v_\rho/2}$ and $\chi = p\eta\rho_\varepsilon\sqrt{2v_\rho^2/V_\rho}$.



Decomposition

$$
\begin{aligned}
&\mathbb{P}_{\theta_c^\star}\Big\{ \bigcup_{n_i \leqslant n < n_{i+1}} E_{c,p}(n,t) \Big\} \\
&\leqslant \quad \mathbb{P}_{\theta_c^\star}\Big\{ \bigcup_{n_i \leqslant n < n_{i+1}} E_{c,p}(n,t) \cap \|\nabla\psi(\theta_c^\star) - \widehat{F}_n\| < \varepsilon_{t,i,c} \Big\} \\
&\qquad + \mathbb{P}_{\theta_c^\star}\Big\{ \bigcup_{n_i \leqslant n < n_{i+1}} E_{c,p}(n,t) \cap \|\nabla\psi(\theta_c^\star) - \widehat{F}_n\| \geqslant \varepsilon_{t,i,c} \Big\} .
\end{aligned}
$$



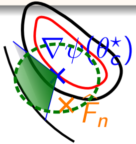
Localization and second change of measure

Localization plus change of measure (first term)

For any sequence of positive values $\{\varepsilon_{t,i,c}\}_{i\geqslant 0}$, it holds

$$\mathbb{P}_{\theta_c^\star}\Big\{\bigcup_{n_i\leqslant n<n_{i+1}} E_{c,p}(n,t)\cap \|\nabla\psi(\widehat{\theta}_n)-\nabla\psi(\theta_c^\star)\| < \varepsilon_{t,i,c}\Big\}$$

$$\leqslant \quad \beta_{\rho,K}\, e^{-f\left(\frac{t}{n_{i+1}-1}\right)} \min\Big\{\rho^2 v_\rho^2, \tilde{\varepsilon}_{t,i,c}^2, \frac{(K+2)v_\rho^2}{K(n_{i+1}-1)V_\rho}\Big\}^{-K/2} \tilde{\varepsilon}_{t,i,c}^K\,,$$



Localization and second change of measure

Localization plus change of measure (first term)

For any sequence of positive values $\{\varepsilon_{t,i,c}\}_{i\geqslant 0}$, it holds

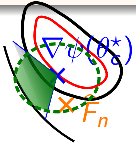
$$\mathbb{P}_{\theta_c^\star}\Big\{\bigcup_{n_i\leqslant n<n_{i+1}} E_{c,p}(n,t)\cap\|\nabla\psi(\widehat{\theta}_n)-\nabla\psi(\theta_c^\star)\|<\varepsilon_{t,i,c}\Big\}$$

$$\leqslant \quad \beta_{\rho,K}e^{-f\left(\frac{t}{n_{i+1}-1}\right)}\min\Big\{\rho^2 v_\rho^2, \tilde{\varepsilon}_{t,i,c}^2, \frac{(K+2)v_\rho^2}{K(n_{i+1}-1)V_\rho}\Big\}^{-K/2}\tilde{\varepsilon}_{t,i,c}^K\,,$$

where $\tilde{\varepsilon}_{t,i,c}=\min\{\varepsilon_{t,i,c},\mathrm{Diam}\big(\nabla\psi(\Theta_\rho)\cap\mathcal{C}_p(\theta_c^\star)\big)\}$ and

$\beta_{\rho,K}=\frac{2}{v_\rho^K}\left(\frac{V_\rho}{v_\rho}\right)^{3K/2}\frac{\omega_{p,K-2}}{\omega_{p',K-2}}$ with $p'>\max\{p,\frac{2}{\sqrt{5}}\}$, with

$\omega_{p,K}=\int_p^1\sqrt{1-z^2}^K dz$ for $K\geqslant 0$ and $w_{p,-1}=1$.



Concentration of measure and boundary effects

We recall that $\nabla\psi(\widehat{\theta}_n) = \widehat{F}_n = \frac{1}{n}\sum_{i=1}^n F(X_i) \in \mathbb{R}^K$, and that $\mathcal{C}_p(\theta_c^\star) = \{\theta \in \Theta : \langle \frac{\theta^\star - \theta_c^\star}{\|\theta^\star - \theta_c^\star\|}, \frac{\nabla\psi(\theta_c^\star) - \nabla\psi(\theta)}{\|\nabla\psi(\theta_c^\star) - \nabla\psi(\theta)\|} \rangle \geqslant p\}$.

Concentration of measure (second term)

Let $\varepsilon_c^{\max} = \mathrm{Diam}(\nabla\psi(\Theta_\rho \cap \mathcal{C}_p(\theta_c^\star)))$. Then, for all $\varepsilon_{t,i,c}$, it holds

$$\mathbb{P}_{\theta_c^\star}\Big\{ \bigcup_{n=n_i}^{n_{i+1}-1} E_{c,p}(n,t) \cap \|\nabla\psi(\widehat{\theta}_n) - \nabla\psi(\theta_c^\star)\| \geqslant \varepsilon_{t,i,c} \Big\}$$

$$\leqslant \exp\left(- \frac{n_i^2 p \varepsilon_{t,i,c}^2}{2 V_\rho (n_{i+1} - 1)} \right) \mathbb{I}\{\varepsilon_{t,i,c} \leqslant \overline{\varepsilon}_c\}.$$

Remark

Non trivial due to the boundary of the space.

Combining the different steps

$$\mathbb{P}_{\theta^\star}\Big\{\bigcup_{1\leqslant n\leqslant t}\widehat{\theta}_n\in\Theta_\rho\cap\mathcal{K}(\Pi(\widehat{\nu}_n),\mu^\star-\varepsilon)\geqslant f(t/n)/n\Big\}\leqslant$$

$$\sum_{c=1}^{C}\sum_{i=0}^{I_t-1}\underbrace{\exp\Big(-n_i\alpha^2-\chi\sqrt{n_if(t/n_i)}\Big)}_{\text{change of measure}}\Big[\underbrace{\exp\Big(-\frac{n_i^2p\varepsilon_{t,i,c}^2}{2V_\rho(n_{i+1}-1)}\Big)\mathbb{I}\{\varepsilon_{t,i,c}\leqslant\overline{\varepsilon}_c\}}_{\text{concentration}}$$

$$+\underbrace{\beta_{p,K}\exp\Big(-f\Big(\frac{t}{n_{i+1}-1}\Big)\Big)\min\Big\{\rho^2v_\rho^2,\varepsilon_{t,i,c}^2,\frac{(K+2)v_\rho^2}{K(n_{i+1}-1)V_\rho}\Big\}^{-K/2}\varepsilon_{t,i,c}^K}_{\text{localization + change of measure}}\Big],$$

Boundary crossing for $f(t)$

- Choose $\varepsilon_{t,i,c} = \sqrt{\frac{2V_\rho(n_{i+1}-1)f(t/(n_{i+1}-1))}{pn_i^2}}$ and $n_i = b^i$:

$$\mathbb{P}_{\theta^\star}\Big\{\bigcup_{1\leqslant n<t} \widehat{\theta}_n \in \Theta_\rho \cap \mathcal{K}(\Pi(\widehat{\nu}_n), \mu^\star - \varepsilon) \geqslant f(t)/n\Big\}$$
$$\leqslant \quad \frac{C}{t}\sum_{i=0}^{I_t-1} \underbrace{e^{-\alpha^2 b^i - \chi\sqrt{b^i f(t)}}}_{s_i} \ln(t)^{K/2-\xi}\Big(1+\xi\frac{\ln\ln(t)}{\ln(t)}\Big)^{K/2}.$$

- idea: Tight control of $\frac{s_{i+1}}{s_i}$.
- This enables to go up to $\xi \gtrsim K/2-1$, instead of $\xi > K/2+1$.
- Similar (but more involved) approach for $f(t/n)$.

Conclusion

A tribute to T.L Lai

Summary

Tribute to T.L. Lai

- 30 years ago: sharp understanding of boundary crossing probabilities (**Read old papers!**)
- Key proof based on **change of measure** argument.
- **Cone** constraint plus sharp peeling.

Modern rewriting

- Non-asymptotic result plus more explicit/smaller constants.
- Complete proof for dimension K.
- Tricky steps: cone covering, cone-constrained concentration inequalities.
- Guarantee for `KL-ucb` and `KL-ucb+` for exponential families of dimension K (out of reach of previous analyses).

Bibliography

T.L. Lai and H. Robbins.
Asymptotically efficient adaptive allocation rules.
Advances in Applied Mathematics, 6(1):4–22, 1985.

Tze Leung Lai.
Adaptive treatment allocation and the multi-armed bandit problem.
The Annals of Statistics, pages 1091–1114, 1987.

Tze Leung Lai.
Boundary crossing problems for sample means.
The Annals of Probability, pages 375–396, 1988.

Bibliography

J. Honda and A. Takemura.
An asymptotically optimal bandit algorithm for bounded support models.
In T. Kalai and M. Mohri, editors, *Conf. Comput. Learning Theory*, Haifa, Israel, 2010.

O. Cappé, A. Garivier, O-A. M., R. Munos, and G. Stoltz.
Kullback–Leibler upper confidence bounds for optimal sequential allocation.
Annals of Statistics, 41(3):1516–1541, 2013.

MERCI

ありがとうございます

Inria Lille - Nord Europe

odalricambrym.maillard@inria.fr

odalricambrymmaillard.wordpress.com