Regret Minimization on Non-Parametric Bandits

Via The Empirical Likelihood Method

Aurélien Garivier[†], joint work with Gilles Stoltz[*], Hédi Hadiji[*] and Pierre Ménard[†]

Séminaire de probabilités de l'UMPA, 26 avril 2018

[†] Équipe-projet AOC: Apprentissage, Optimisation, Complexité
Institut de Mathématiques de Toulouse LabeX CIMI Université Paul Sabatier Toulouse III
[*] Laboratoire de Mathmatiques, CNRS Equipe de Probabilités, Statistique et Modélisation
Université Paris-Sud

Table of contents

Model

Simple Non-parametric Bandits

K arms $\underline{\nu} = (\nu_a)_{1 \le a \le K}$, $\nu_a \in \mathcal{P}[0,1]$

$\mu_a = \mathrm{E}(\nu_a)$, $\quad \mu^\star = \mu_{a^\star} = \max_a \mu_a < 1$

Random observations $(X_{a,n})_{1 \le a \le K, n \ge 1}$ independent, $X_{a,n} \sim \nu_a$.

Bandit Setting

At each round $t = 1, 2, \dots, T$:

- Choose $A_t = \phi_t\big(I_{t-1}\big)$ where $I_{t-1} = (A_s, Y_s)_{1 \le s < t}$
- Observe $Y_t = X_{A_t, N_{A_t}(t)}$ independent sample of ν_{A_t} .

where for all $t \ge 1$

$$N_a(t) = \sum_{s \le t} \mathbb{1}\{A_s = a\} \, .$$

Bandit Observations

n=	1	2	3	4	5	6	7	8	9
Arm 1	0.55	0.98	0.13	0.24	0.44	0.62	0.69	0.88	0.75
Arm 2	0.21	0.17	0.03	0.24	0.67	0.25	0.72	0.53	0.71
Arm 3	0.37	0.17	0.16	0.54	1.00	0.92	0.41	0.93	0.16
Arm 4	0.05	0.76	0.22	0.79	0.85	0.86	0.35	0.34	0.86

t= 0

$$\hat{\nu}_a(t) = \frac{1}{N_a(t)} \sum_{n \leq N_a(t)} \delta_{X_{a,n}}$$

Bandit Observations

n=	1	2	3	4	5	6	7	8	9	
Arm 1	0.55	0.98	0.13	0.24	0.44	0.62	0.69	0.88	0.75	N1(30)=8
Arm 2	0.21	0.17	0.03	0.24	0.67	0.25	0.72	0.53		N2(30)=8
Arm 3	0.37	0.17	0.16	0.54	1.00	0.92	0.41			N3(30)=7
Arm 4	0.05	0.76	0.22	0.79	0.85	0.86				N4(30)=6

t=30

$$\hat{\nu}_a(t) = \frac{1}{N_a(t)} \sum_{n \leq N_a(t)} \delta_{X_{a,n}}$$

Regret Minimization

Goal

Find a strategy $\underline{\phi} = (\phi_t)_{t\geq 1}$ so as to maximize $\sum_{t=1}^{T} Y_t$ in expectation.

Equivalent to minimizing the

Expected Regret

$$R_T(\underline{\phi}, \underline{\nu}) = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right]$$

$$= \sum_{a=1}^{K} (\mu^\star - \mu_a)\, \mathbb{E}\big[N_a(T)\big] .$$

The strategy $\underline{\phi}$ must minimize $\mathbb{E}\big[N_a(T)\big]$ for all a such that $\mu_a < \mu^\star$.

Lower Bound

Lower Bound

See Lai and Robbins [1985], Burnetas and Katehakis [1996], Garivier et al. [2018]

It the strategy $\underline{\phi}$ is such that such that for all bandit problems $\underline{\nu}$ over $[0,1]$, for all suboptimal arms a,

$$\forall\, \alpha > 0, \qquad \mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] = o(T^{\alpha})\,,$$

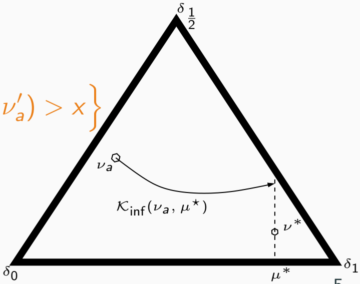
then for any bandit problem $\underline{\nu}$ over $[0,1]$, for any suboptimal arm a,

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]}{\ln T} \geq \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\,.$$

where for all $x \in [0,1]$,

$$\mathcal{K}_{\inf}(\nu_a, x) = \inf\Big\{\mathrm{KL}(\nu_a, \nu'_a) :\ \nu'_a \in \mathcal{P}[0,1]\ ,\ \mathrm{E}(\nu'_a) > x\Big\}$$

and where KL denotes the Kullback-Leibler divergence.



Lower Bound: Sktech of Proof

Let $I_t = \big(Y_1, A_1, \ldots, Y_t, A_t\big)$ be the variables observed up to time t.
Then, for every $\underline{\nu}' = \big(\nu'_a\big)_{1\le a\le K}$,

$$\sum_{a=1}^{K} \mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]\,\mathrm{KL}(\nu_a,\nu'_a) = \mathrm{KL}\big(\mathbb{P}_{\underline{\nu}}^{I_T},\, \mathbb{P}_{\underline{\nu}'}^{I_T}\big)$$

$$\ge \mathrm{KL}\Big(\mathbb{P}_{\underline{\nu}}^{N_a(T)},\, \mathbb{P}_{\underline{\nu}'}^{N_a(T)}\Big)$$

$$\ge \mathrm{kl}\left(\mathbb{E}_{\underline{\nu}}\left[\frac{N_a(T)}{T}\right],\, \mathbb{E}_{\underline{\nu}'}\left[\frac{N_a(T)}{T}\right]\right)$$

where $\mathrm{kl}(x,y) = x\ln\frac{x}{y} + (1-x)\ln\frac{1-x}{1-y}$ is the binary relative entropy.

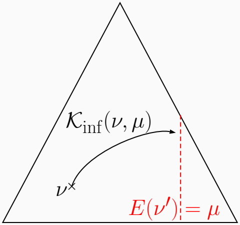
If $\nu'_b = \nu_b$ for all $b \neq a$, and if $\mathrm{E}(\nu'_a) > \mu^\star$, then $\forall \alpha > 0$

$$\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]\,\mathrm{KL}\big(\nu_a,\nu'_a\big) \ge \mathrm{kl}\left(\frac{o(T^\alpha)}{T},\, 1-\frac{o(T^\alpha)}{T}\right) \sim (1-\alpha)\ln(T)$$

and

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]}{\ln T} \ge \frac{1}{\mathcal{K}_{\inf}(\nu_a,\mu^\star)}\,.$$

Regularity of $\mathcal{K}_{\inf}$



$$\mathcal{K}_{\inf}(\nu, \mu) \underbrace{\geq \mathrm{kl}\big(\,\mathrm{E}[\nu], \mu\big)}_{\textit{contraction}} \underbrace{\geq 2(\mathrm{E}[\nu] - \mu)^2}_{\textit{Pinsker}}$$

Proposition

For all $\nu \in \mathcal{P}[0,1]$, all $\mu \in (0,1)$, and all $\varepsilon \in (0, \min\{\mu, \mu - \mathrm{E}(\nu)\})$,

$$2\varepsilon^2 \leq \mathcal{K}_{\inf}(\nu, \mu) - \mathcal{K}_{\inf}(\nu, \mu - \varepsilon) \leq \frac{\varepsilon}{1-\mu}\,.$$

Computing $\mathcal{K}_{\text{inf}}$

Variational Formula

For all $\nu \in \mathcal{P}[0,1]$ and all $0 < \mu < 1$,

$$\mathcal{K}_{\text{inf}}(\nu, \mu) = \max_{0 \le \lambda \le 1} \int_0^1 \ln\left(1 - \lambda \frac{x-\mu}{1-\mu}\right) d\nu(x)$$

Moreover, if we denote by $\lambda^\star$ the value at which the above maximum is reached, then $\mathcal{K}_{\text{inf}}(\nu, \mu) = \mathrm{KL}(\nu, \tilde{\nu}_\mu)$ where

$$d\tilde{\nu}_\mu(x) = \frac{d\nu(x)}{1 - \lambda^\star \frac{x-\mu}{1-\mu}} + r\, d\delta_1(x)$$

with $r = 1 - \int_0^1 \frac{d\nu(x)}{1-\lambda^\star(x-\mu)/(1-\mu)} \in [0,1)$.

Proof

- (boring) Check that there exists $\lambda^\star \in [0,1]$ and $r \geq 0$ such that

$$d\tilde{\nu}_\mu(x) = \frac{d\nu(x)}{1 - \lambda^\star \frac{x-\mu}{1-\mu}} + r\, d\delta_1(x)$$

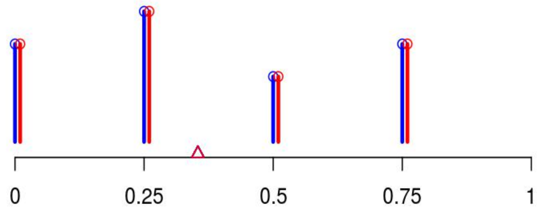
is a probability measure with $E(\tilde{\nu}_\mu) = \mu$, and $\lambda^\star = 1 \implies \nu\big(\{1\}\big) = 0$.

$$\mathcal{K}_{\text{inf}}(\nu,\mu) \geq \mathrm{KL}(\nu,\tilde{\nu}_\mu) = \int_0^1 \ln\left(1 - \lambda^\star \frac{x-\mu}{1-\mu}\right) d\nu(x)$$

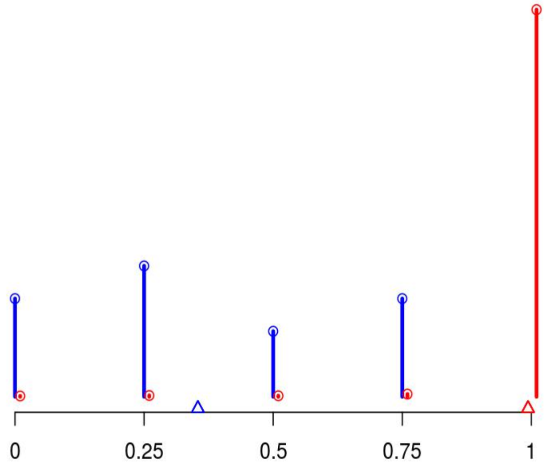
- Let $\nu' \in \mathcal{P}[0,1]$ be such that $E(\nu') \geq \mu$, and $\nu' \gg \nu$. Then

$$\begin{aligned}
\mathrm{KL}(\nu,\nu') - \mathrm{KL}(\nu,\tilde{\nu}_\mu) &= -\int_0^1 \ln\left(\frac{\frac{d\nu}{d\tilde{\nu}_\mu}(x)}{\frac{d\nu}{d\nu'}(x)}\right) d\nu(x) \\
&\geq -\ln\left[\int_0^1 \frac{\frac{d\nu}{d\tilde{\nu}_\mu}(x)}{\frac{d\nu}{d\nu'}(x)}\, d\nu(x)\right] \\
&\geq -\ln\left[\int_0^1 \left(1 - \lambda^\star \frac{x-\mu}{1-\mu}\right) d\nu'(x)\right] \\
&\geq -\ln(1)\,.
\end{aligned}$$

Closest Distribution with Mean μ



Closest Distribution with Mean μ



The KL-UCB strategy

Optimism in face of Uncertainty

Upper Confidence Bound (UCB) Strategies

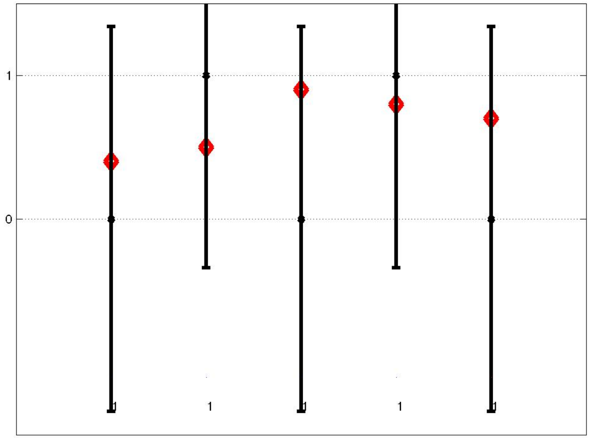
At each round $t = 1, 2, \dots, T$:

- Compute an UCB $U_a(t)$ for all $a \in \{1, \dots, K\}$
- Choose $A_t = \text{argmax}_a U_a(t)$

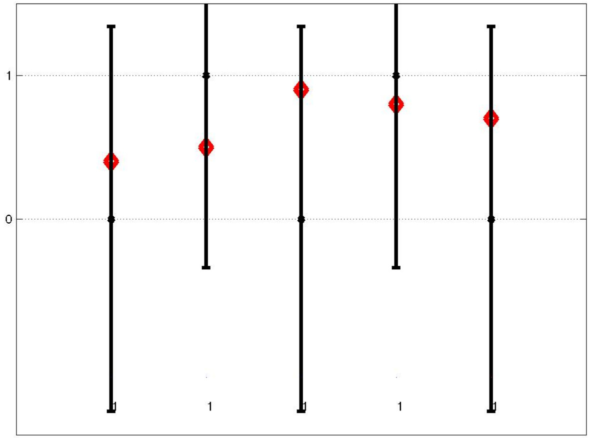
Remark: in Bayesian parametric bandits, the optimal policy is an *index policy* that, in some asymptotics, mimics UCB.

See Gittins [1979], Chang and Lai [1987].

UCB in Action



UCB in Action

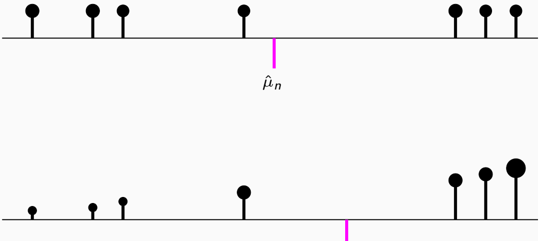


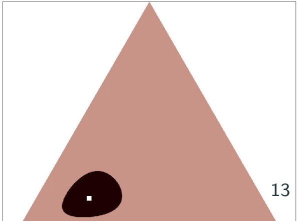
Non-parametric confidence bounds: Empirical Likelihood Owen [2001]

Empirical measure:

$$\hat{\nu}_a(t) = \frac{1}{N_a(t)} \sum_{s \le t} \delta_{Y_s} \mathbb{1}\{A_s = a\}$$

$$\begin{aligned} U_a(t) &= \sup\left\{ \mathrm{E}\left(\nu'\right) \;\middle|\; \nu' \in \mathcal{P}[0,1], \mathrm{KL}\left(\hat{\nu}_a(t), \nu'\right) \le \frac{\ln(T)}{N_a(t)} \right\} \\ &= \sup\left\{ \mu \in [0,1] \;\middle|\; \mathcal{K}_{\inf}\left(\hat{\nu}_a(t), \mu\right) \le \frac{\ln(T)}{N_a(t)} \right\} \end{aligned}$$



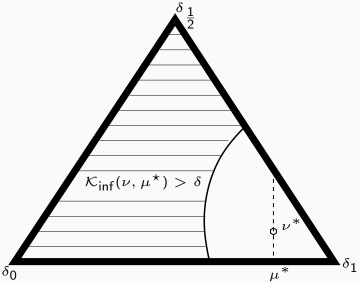


Calibrating EL Confidence Bounds

Deviation bound on $\mathcal{K}_{\text{inf}}$

Let $\hat{\nu}_n$ denote the empirical distribution associated with a sequence of n i.i.d. random variables with distribution ν over $[0,1]$ with $\mathrm{E}(\nu) \in (0,1)$. Then, for all $u \geq 0$,

$$\mathbb{P}\Big[\mathcal{K}_{\text{inf}}\big(\hat{\nu}_n, \mathrm{E}(\nu)\big) \geq u\Big] \leq e(2n+1)\, e^{-nu}$$



Proof of the deviation bound

$$\mathcal{K}_{\inf}\big(\hat{\nu}_n, \mathrm{E}(\nu)\big) = \max_{0 \le \lambda \le 1} G(\lambda) \qquad \text{where} \quad G(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 - \lambda \frac{X_i - \mu}{1-\mu}\right) \; .$$

Prop: for $\epsilon > 0$, let

$$\Lambda_\epsilon = \left\{ \frac{1}{2} - \left\lfloor \frac{1}{2\epsilon} \right\rfloor \epsilon, \ldots, \frac{1}{2} - \epsilon, \frac{1}{2}, \frac{1}{2} + \epsilon, \ldots, \frac{1}{2} - \left\lfloor \frac{1}{2\epsilon} \right\rfloor \epsilon \right\} \; .$$

Then $\left|\Lambda_\epsilon\right| \le 1 + 1/\epsilon$ and

$$\forall \lambda \in [0,1], \exists \lambda' \in \Lambda_\epsilon : G(\lambda') \ge G(\lambda) - 2\epsilon \; .$$

Proof of the deviation bound

$$\mathcal{K}_{\text{inf}}\big(\hat{\nu}_n, \mathrm{E}(\nu)\big) = \max_{0\le\lambda\le1} G(\lambda) \qquad \text{where} \quad G(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 - \lambda\frac{X_i - \mu}{1-\mu}\right) .$$

$$\begin{aligned}
\mathbb{P}\Big[\mathcal{K}_{\text{inf}}\big(\hat{\nu}_n, \mathrm{E}(\nu)\big) \ge u\Big] &\le \mathbb{P}\Big[\max_{\lambda\in\Lambda_\epsilon} G(\lambda) \ge u - 2\epsilon\Big] \\
&\le \sum_{\lambda\in\Lambda_\epsilon} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\ln\left(1-\lambda\frac{X_i-\mu}{1-\mu}\right) \ge u - 2\epsilon\right) \\
&\le \sum_{\lambda\in\Lambda_\epsilon} \mathbb{E}\left[\prod_{i=1}^{n}\left(1-\lambda\frac{X_i-\mu}{1-\mu}\right)\right] e^{-n(u-2\epsilon)} \\
&\le \big|\Lambda_\epsilon\big| e^{-n(u-2\epsilon)} \\
&= (2n+1)e^{-nu+1}
\end{aligned}$$

for $\epsilon = 1/(2n)$.

A Vanilla Regret Analysis

Regret Bound

See [Garivier, Hadiji, Ménard, Stoltz 2018], see also Cappé et al. [2013], Honda and Takemura [2015] and references therein

Theorem

For all arms a such that $\mu_a < \mu^\star$, the KL-UCB strategy ensures that

$$\mathbb{E}_{\underline{\nu}}\big[N_a(t)\big] \leq \frac{\ln(T)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\big(1 + o(1)\big) .$$

Thus, the KL-UCB strategy is optimal in the long run.

Decomposition of the Regret

Let $\delta > 0$ to be chosen later ($\delta = 1/\ln(T)^{1/5}$)

$$\mathbb{E}_{\underline{\nu}}\left[N_a(T)\right] = 1 + \sum_{t=K}^{T-1} \mathbb{P}_{\underline{\nu}}(U_a(t) < \mu^\star - \delta, A_{t+1=a}) \tag{1}$$

$$+ \sum_{t=K}^{T-1} \mathbb{P}_{\underline{\nu}}(U_a(t) \geq \mu^\star - \delta, A_{t+1=a}) \tag{2}$$

(1) $\rightarrow$ underestimation of the optimal arm

(2) $\rightarrow$ normal when $N_a(t)$ is small, otherwise overestimation of arm a

Upper-bounding Term (1) - 1/2

Since $U_{A_t}(t) \geq U_{a^\star}(t)$ for every $t \geq K$,

$$
\begin{aligned}
\left\{U_a(t) < \mu^\star - \delta, A_{t+1} = a\right\} &\subset \left\{\exists t \leq T : U_{a^\star}(t) < \mu^\star - \delta\right\} \\
&\subset \left\{\exists n \leq T : U_{a^\star,n} < \mu^\star - \delta\right\} \\
&\subset \left\{\exists n \leq T : \mathcal{K}_{\inf}\left(\hat{\nu}_{a^\star,n}, \mu^\star - \delta\right) > \frac{\ln(T)}{n}\right\} \ .
\end{aligned}
$$

Upper-bounding Term (1) - 2/2

$$
\begin{aligned}
\mathbb{P}_{\underline{\nu}}\Big(\exists n \le T : \mathcal{K}_{\inf}\big(\hat{\nu}_{a^\star,n}, \mu^\star - \delta\big) > \frac{\ln(T)}{n}\Big)
&\le \sum_{n=1}^{T} \mathbb{P}_{\underline{\nu}}\left(\mathcal{K}_{\inf}\big(\hat{\nu}_{a^\star,n}, \mu^\star - \delta\big) > \frac{\ln(T)}{n}\right) \\
&\le \sum_{n=1}^{\infty} \mathbb{P}_{\underline{\nu}}\left(\mathcal{K}_{\inf}\big(\hat{\nu}_{a^\star,n}, \mu^\star\big) > 2\delta^2 + \frac{\ln(T)}{n}\right) \\
&\le \sum_{n=1}^{\infty} \frac{e(2n+1)}{T}\, e^{-2n\delta^2} \\
&\le \frac{C}{T\delta^4} \qquad \text{since } \sum_{n=1}^{\infty} n e^{-n\theta} = \frac{e^{-\theta}}{\left(1-e^{-\theta}\right)^2} \le \left(1+\frac{1}{\theta}\right)^2
\end{aligned}
$$

and hence

$$
\sum_{t=K}^{T-1} \mathbb{P}_{\underline{\nu}}(U_a(t) < \mu^\star - \delta, A_{t+1=a}) \le \frac{C}{\delta^4}\,.
$$

Upper-bounding Term (2) - 1/2

Let

$$
\begin{aligned}
n_0 &= \min\left\{n : \mathcal{K}_{\inf}(\nu_a, \mu^\star - 2\delta) \le \frac{\ln(T)}{n}\right\} \\
&= \left\lceil \frac{\ln(T)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star - 2\delta)} \right\rceil \\
&\le 1 + \frac{\ln(T)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star) - 4\delta^2}
\end{aligned}
$$

Then

$$
\begin{aligned}
&\sum_{t=K}^{T-1} \mathbb{P}_{\underline{\nu}}\big(U_a(t) \ge \mu^\star - \delta, A_{t+1=a}\big) \\
&\qquad \le n_0 + \sum_{t=n_0}^{T-1} \mathbb{P}_{\underline{\nu}}\big(U_a(t) \ge \mu^\star - 2\delta, A_{t+1=a}, N_a(t) > n_0\big) \\
&\qquad \le n_0 + \sum_{n=n_0}^{T-1} \mathbb{P}_{\underline{\nu}}\big(U_{a,n} \ge \mu^\star - 2\delta\big)
\end{aligned}
$$

Upper-bounding Term (2) - 2/2

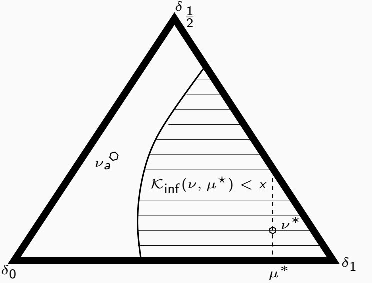
For $n \geq n_0$,

$$\{U_{a,n} \geq \mu^\star - \delta\} \subset \left\{\mathcal{K}_{\inf}(\hat{\nu}_{a,n}, \mu^\star) \leq \frac{\ln(T)}{n}\right\}$$
$$\subset \{\mathcal{K}_{\inf}(\hat{\nu}_{a,n}, \mu^\star - \delta) \leq \mathcal{K}_{\inf}(\nu_a, \mu^\star - 2\delta)\}$$
$$\subset \left\{\mathcal{K}_{\inf}(\hat{\nu}_{a,n}, \mu^\star - \delta) \leq \mathcal{K}_{\inf}(\nu_a, \mu^\star - \delta) - \frac{\delta}{1-\mu^\star}\right\}$$

and $\mathbb{P}\big(U_{a,n} \geq \mu^\star - \delta\big) \leq \exp\big(-\epsilon(\delta)n\big)$

with

$$\epsilon(\delta) = \inf\left\{KL(\nu, \nu_a) : \mathcal{K}_{\inf}(\nu, \mu^\star - \delta) < \mathcal{K}_{\inf}(\nu_a, \mu^\star - \delta) - \frac{\delta}{1-\mu^\star}\right\}$$
$$\geq \varepsilon\, \delta^2 .$$



Collecting Everything

$$\begin{aligned}
\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] &= 1 + \sum_{t=K}^{T-1} \mathbb{P}_{\underline{\nu}}(U_a(t) < \mu^\star - \delta, A_{t+1=a}) \\
&\qquad + \sum_{t=K}^{T-1} \mathbb{P}_{\underline{\nu}}(U_a(t) \geq \mu^\star - \delta, A_{t+1=a}) \\
&\leq 1 + \frac{C}{\delta^4} + 1 + \frac{\ln(T)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star) - 4\delta^2} + \frac{1}{\epsilon\delta^2} \\
&= \frac{\ln(T)}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)}\big(1 + o(1)\big)
\end{aligned}$$

for $\delta = 1/\ln(T)^{1/5}$.

Results and Questions

Careful Lower bound

Theorem Garivier et al. [2018]

For uniformly super-fast convergent strategies, that is, strategies for which there exists a constant C such for all bandit problems $\underline{\nu}$ over $[0,1]$, for all suboptimal arms a,

$$\frac{\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]}{\ln T} \leq \frac{C}{\Delta_a^2},$$

the lower bound above can be strengthened into: for any bandit problem $\underline{\nu}$ over $[0,1]$, for any suboptimal arm a,

$$\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] \geq \frac{\ln T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} - \Omega\big(\ln(\ln T)\big).$$

Initial Regime

Theorem Garivier et al. [2018]

We say that a strategy ψ is *smarter than uniform* for all bandit problems $\underline{\nu}$, for all optimal arms $a^\star$, for all $T \geq 1$,

$$\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a^\star}(T)\big] \geq \frac{T}{K}\,.$$

For all strategies ψ that are smarter than uniform, for all bandit problems $\underline{\nu}$, for all arms a, for all $T \geq 1$,

$$\mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \geq \frac{T}{K}\Big(1 - \sqrt{2T\mathcal{K}_{\inf}(\nu_a,\mu^\star)}\Big)\,.$$

In particular,

$$\forall\, T \leq \frac{1}{8\mathcal{K}_{\inf}(\nu_a,\mu^\star)}, \qquad \mathbb{E}_{\underline{\nu}}\big[N_{\psi,a}(T)\big] \geq \frac{T}{2K}\,.$$

Minimax Regret

Minimax Lower Bound

For all $T \geq 1$ and all $K \geq 2$,

$$\inf_{\underline{\phi}} \sup_{\underline{\nu}} R_T(\underline{\phi}, \nu) \geq \frac{1}{20} \min\left\{\sqrt{KT},\ T\right\}. \tag{3}$$

Minimax optimal strategy: MOSS Audibert and Bubeck [2009] (but not asymptotically optimal). New analysis in [Garivier, Hadiji, Ménard, Stoltz 2018].

Minimax and Asymptotically Optimal strategies

KL-UCB improved

At each round $t = 1, 2, \dots, T$:

- Compute an UCB $U_a(t)$ for all $a \in \{1, \dots, K\}$
- Choose $A_t = \mathrm{argmax}_a \sup\left\{ \mu \in [0,1] \,\middle|\, \mathcal{K}_{\inf}\big(\hat{\nu}_a(t), \mu\big) \leq \frac{\ln\left(\frac{T}{K\, N_a(t)}\right)}{N_a(t)} \right\}$

Technical (?) complication in KL-UCB-switch

Further complications to make to make it anytime (unaware of the final horizon T)

Non-asymptotic regret bounds

Regret Bounds for KL-UCB-switch [Garivier, Hadiji, Ménard, Stoltz 2018]

Theorem: Distribution-free bound

Given $T \geq 1$, the regret of the KL-UCB-switch algorithm, tuned with the knowledge of T, is uniformly bounded over all bandit problems $\underline{\nu}$ over $[0, 1]$ by

$$R_T \leq (K - 1) + 25\sqrt{KT} \,.$$

Theorem: Distribution-dependent bound

Given $T \geq 1$, the KL-UCB-switch algorithm ensures that for all bandit problems $\underline{\nu}$ over $[0, 1]$, for all sub-optimal arms a,

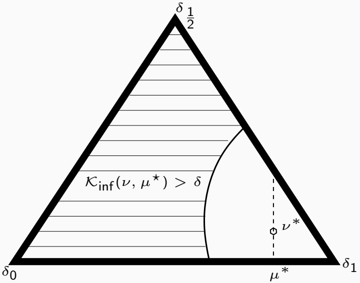
$$\mathbb{E}_{\underline{\nu}}[N_a(T)] \leq \frac{\ln T - \ln \ln T}{\mathcal{K}_{\inf}(\nu_a, \mu^\star)} + O_T(1) \,.$$

Wanted: A better Deviation Bound for $\mathcal{K}_{\text{inf}}$?

Deviation bound on $\mathcal{K}_{\text{inf}}$

Let $\hat{\nu}_n$ denote the empirical distribution associated with a sequence of n i.i.d. random variables with distribution ν over $[0,1]$ with $\mathrm{E}(\nu) \in (0,1)$. Then, for all $u \geq 0$,

$$\mathbb{P}\Big[\mathcal{K}_{\text{inf}}\big(\hat{\nu}_n, \mathrm{E}(\nu)\big) \geq u\Big] \leq e(2n+1)\, e^{-nu} \ .$$



Thank you for your attention!

References

R. Agrawal. Sample mean based index policies with $o(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.

J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory*, COLT'09, pages 217–226, 2009.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

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

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

Fu Chang and Tze Leung Lai. Optimal stopping and dynamic allocation. *Advances in Applied Probability*, 19(4):829–853, 1987. ISSN 00018678. URL `http://www.jstor.org/stable/1427104`.

R. Degenne and V. Perchet. Anytime optimal algorithms in stochastic multi-armed bandits. In *Proceedings of the 2016 International Conference on Machine Learning*, ICML'16, pages 1587–1595, 2016.

A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Annual Conference on Learning Theory*, COLT'11, 2011.

A. Garivier, P. Ménard, and G. Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 2018. To appear; meanwhile, see arXiv preprint arXiv:1602.07182.

J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177, 1979. ISSN 00359246. URL `http://www.jstor.org/stable/2985029`.

J. Honda and A. Takemura. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *Journal of Machine Learning Research*, 16:3721–3756, 2015.

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

O.-A. Maillard, R. Munos, and G. Stoltz. Finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. In *Proceedings of the 24th annual Conference on Learning Theory*, COLT'11, 2011.

P. Ménard and A. Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In *Procedings of the 2017 Algorithmic Learning Theory Conference*, ALT'17, 2017.

A.B. Owen. *Empirical Likelihood*. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, 2001. ISBN 9781584880714. URL http://books.google.fr/books?id=QUDPCpc1TNMC.

W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.