Part 1: Stochastic bandits, setting, upper bounds and statement of the lower bounds

Setting (finitely many arms): K arms indexed by $a = 1, 2, \ldots K$

a given "bandit problem" $\left\{ \begin{array}{l} K \text{ probability distributions } \nu_1, \ldots \nu_K \text{ over } \mathbb{R} \\ \text{with expectations } \mu_1 \ldots \mu_K \end{array} \right.$

$\hookrightarrow$ let $\mu^* = \max\limits_{a=1..K} \mu_a$

These distributions are unknown to the decision-maker.
She will need to perform a trade-off between
- exploration, ie, pulling each arm sufficiently often to estimate its performance
- exploitation, ie, pulling more often the best-performing arms

At each round $t = 1, 2, \ldots$:
1. The decision-maker picks an arm $I_t \in \{1..K\}$, possibly at random (thanks to an auxiliary randomization U_{t-1})
2. She gets a reward Y_t drawn at random according to ν_{I_t} (given I_t)
3. This is the only feedback / piece of information she gets

Definition: A (randomized) bandit strategy ψ is a sequence of measurable functions $(\psi_t)_{t \geq 0}$ with

$$\psi_t : \underbrace{H_t = (U_0, Y_1, U_1, \ldots Y_t, U_t)}_{\text{history for the first } t \text{ rounds}} \longmapsto \underbrace{\psi_t(H_t) = I_{t+1}}_{\text{arm picked at round } t+1}$$

[We will take $U_0, U_1, \ldots$ iid $\sim \mathcal{U}_{[0,1]}$ and independent from all other random variables]

Aim: Control from above ("minimize") the pseudo-regret:

$$R_T = T \max_{a=1..K} \mu_a - \mathbb{E}\Big[\sum_{t=1}^{T} Y_t\Big] = T\mu^* - \mathbb{E}\Big[\sum_{t=1}^{T} Y_t\Big]$$

Comment: it's a very weak notion of regret (very "expected")

Rewriting: Let's start to do some maths!

Since $Y_t \mid I_t \sim \nu_{I_t}$, we have (by the bandit model): $\mathbb{E}[Y_t \mid I_t] = \mu_{I_t}$.

Thus by the tower rule: $\mathbb{E}[Y_t] = \mathbb{E}[\mu_{I_t}] = \sum_{a=1}^{K} \mu_a \, \mathbb{P}\{I_t = a\}$

Denote $N_a(T) = \sum_{t=1}^{T} \mathbb{1}_{\{I_t = a\}}$ and note that $\sum_{a=1}^{K} N_a(T) = T$

Thus
$$R_T = \mu^* \sum_{a=1}^{K} \mathbb{E}[N_a(T)] - \sum_{a=1}^{K} \mu_a \sum_{t=1}^{K} \mathbb{P}\{I_t = a\}$$
$$= \sum_{a=1}^{K} \underbrace{(\mu^* - \mu_a)}_{= \Delta_a, \text{ the gap of arm } a} \mathbb{E}[N_a(T)] = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)]$$

Conclusion : to upper or lower bound the regret, it is necessary and sufficient to upper or lower bound the $\mathbb{E}[N_a(T)]$ for arms a with $\Delta_a > 0$ (called: suboptimal arms).

<u>Typical upper bounds</u> :

1. <u>Distribution-free bounds</u> : when the ν_a are distributions over a known compact set; e.g. over $[0,1]$

The MOSS strategy by Audibert and Bubeck achieves: $\forall T \geqslant 1$,

$$\sup_{\substack{\nu_1 \ldots \nu_K \\ \text{proba. distr. over} \\ [0,1]}} R_T \leqslant 49\sqrt{TK}$$

2. <u>Distribution-dependent bounds</u> : We call a bandit model $\mathcal{D}$ a collection of possible probability distributions, each having an expectation.

E.g., in the example above, we considered: $\mathcal{D} = \mathcal{P}([0,1])$, the set of all probability distributions over $[0,1]$

For "well-behaved" models $\mathcal{D}$ (I don't want to make this statement more precise!), the following bound can be achieved, in an asymptotic or non-asymptotic way:

For all bandit problems $(\nu_a)_{a=1..K}$ in $\mathcal{D}$,

$$\limsup_{T \to +\infty} \frac{\mathbb{E}[N_a(T)]}{\ln T} \leqslant \frac{1}{K_{\inf}(\nu_a, \mu^*, \mathcal{D})}$$

where $K_{\inf}(\nu_a, \mu^*, \mathcal{D}) = \inf \left\{ KL(\nu_a, \nu_a') : \nu_a' \in \mathcal{D} \text{ and } E(\nu_a') > \mu^* \right\}$

Kullback-Leibler divergence, I will recall it tomorrow!

expectation of ν_a'

<u>Ex</u>: $\mathcal{D} = \{ Ber(p), p \in [0,1] \}$: Lai and Robbins, 1985

$\mathcal{D}$ = any family indexed in a smooth way by a finite-dimension parameter: Burnetas and Katehakis, 1996

$\mathcal{D}$ = 1-dimensional regular exponential family (with a non-asymptotic bound): Garivier et Cappé, 2011

$\mathcal{D} = \mathcal{P}([0,1])$: Honda and Takemura, 2015

<u>Lower bounds</u> : The aim of this short course!

1. <u>Distribution-free bound of the order of $\sqrt{TK}$</u> :

Perfectly provided by Auer, Cesa-Bianchi, Freund and Schapire, 1995, 2002

But I will recall its proof!

Theorem 2: For all $K \geqslant 2$ and $T \geqslant K/5$,

$$\inf_{\substack{\text{strategies}\\ \psi}} \; \sup_{\substack{\nu_1 \ldots \nu_K \\ \text{in } \mathcal{P}([0,1])}} R_T \geqslant \frac{1}{20}\sqrt{TK}$$

2. Distribution-dependent bound: Since we want a bound for each bandit problem considered individually, we need to rule out strategies that are not "reasonable" (that would perform very well on some bandit problems at the cost of performing very badly on others!).

Definition: A strategy ψ is uniformly fast convergent on a model $\mathcal{D}$ if

for all bandit problems $(\nu_a)_{a=1\ldots K}$ in $\mathcal{D}$,
for all suboptimal arms a (ie, with $\Delta_a > 0$),
for all $\alpha \in (0,1]$,

$$\mathbb{E}[N_a(T)] = o(T^\alpha).$$

As indicated above, for many models $\mathcal{D}$, there exist strategies with $\mathbb{E}[N_a(T)] = O(\ln T) = o(T^\alpha)$ for all $\alpha \in (0,1]$.

Theorem 1: For all bandit models $\mathcal{D}$,
for all strategies ψ that are uniformly fast convergent on $\mathcal{D}$,
for all bandit problems $(\nu_a)_{a=1\ldots K}$ in $\mathcal{D}$,
for all suboptimal arms a (ie, with $\Delta_a > 0$),

$$\liminf_{T \to +\infty} \frac{\mathbb{E}[N_a(T)]}{\ln T} \geqslant \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^*, \mathcal{D})}$$

References: Basically, a widely believed result since the lower bounds by Lai and Robbins, 1985 and Burnetas and Katehakis, 1996 for the models considered therein; see also Cowan and Katehakis, 2015

But all these proofs required some (mild?) assumptions on $\mathcal{D}$.

As a side / warm-up result, we (Garivier, Ménard, Stoltz 2017) obtained this distribution-dependent lower bound without any assumption on $\mathcal{D}$. We hadn't really realized that till the reviewers commented on this (the focus of our article was to obtain linear lower bounds on the regret for small T).

Now that we stated the results, let me give you an overview of the proof techniques (ie, the program for tomorrow!).

The key idea is: we will perform IMPLICIT changes of measures.

Sketch of the proof / Main ideas behind the proof of the distribution-dependent bound

(1) Implicit change of measure: fix $\mathcal{D}$, ψ, $(\nu_k)_{k=1..K}$ and a with $\Delta_a > 0$

Consider the bandit problem $(\nu'_k)_{k=1..K}$ defined as

$$\begin{cases} \nu'_a \text{ is any distribution such that } E(\nu'_a) > \mu^* \text{ and } \nu'_a \in \mathcal{D} \\ \nu'_k = \nu_k \quad \text{for all } k \neq a \end{cases}$$

We consider these alternative problems $(\nu'_k)_{k=1..K}$ because of the defining infimum of $K_{\inf}(\nu_a, \mu^*, \mathcal{D})$.

To avoid any ambiguity, we write $\underline{\nu} = (\nu_k)_{k=1..K}$ and $\underline{\nu}' = (\nu'_k)_{k=1..K}$ and index all probabilities and expectations by the underlying bandit problem.

(2) We establish a fundamental inequality, valid for all pairs of bandit problems:

for all random variables Z taking values in $[0,1]$ and being $\sigma(H_T)$-measurable

$$\sum_{k=1}^{K} \mathbb{E}_{\underline{\nu}}[N_k(T)] \, KL(\nu_k, \nu'_k) = KL\left(\mathbb{P}_{\underline{\nu}}^{H_T}, \mathbb{P}_{\underline{\nu}'}^{H_T}\right) \geq kl\left(\mathbb{E}_{\underline{\nu}}[Z], \mathbb{E}_{\underline{\nu}'}[Z]\right)$$

where $\begin{cases} \mathbb{P}_{\underline{\nu}}^{H_T} \text{ is the law of } H_T \text{ when the bandit problem is } \underline{\nu} \\ kl(p,q) = KL(\mathrm{Ber}(p), \mathrm{Ber}(q)) = p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q} \end{cases}$

Ingredients of the proof: $\begin{cases} = \text{ is obtained by the chain rule for KL} \\ \geq \text{ corresponds to a data-processing inequality with expectations of random variables} \end{cases}$

(I will do a 1h-

(3) Application of (2) on (1): $Z = N_a(T)/T \in [0,1]$ and is $\sigma(H_T)$-measurable

$$\mathbb{E}_{\underline{\nu}}[N_a(T)] \, KL(\nu_a, \nu'_a) \geq kl\left(\mathbb{E}_{\underline{\nu}}[N_a(T)/T], \mathbb{E}_{\underline{\nu}'}[N_a(T)/T]\right)$$
$$\geq \underbrace{\left(1 - \mathbb{E}_{\underline{\nu}}[N_a(T)/T]\right)}_{\to 0} \ln \frac{T}{\underbrace{T - \mathbb{E}_{\underline{\nu}'}[N_a(T)]}} - \ln 2$$

as $kl(p,q) \geq (1-p) \ln \frac{1}{1-q} - \ln 2$

$\to 0$ by uniform fast convergence + a suboptimal for $\underline{\nu}$

$T - \mathbb{E}_{\underline{\nu}'}[N_a(T)] = o(T^{\alpha})$ as a is the unique optimal arm for $\underline{\nu}'$

$\hookrightarrow$ for all $\alpha \in (0,1]$

That is, we have:

$$\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] \; \mathrm{KL}(\nu_a, \nu'_a) \geqslant \big(1-o(1)\big) \underbrace{\ln \frac{T}{o(T^\alpha)}}_{\geqslant \ln T/T^\alpha \,=\, (1-\alpha)\ln T \text{ for } T \text{ large enough}} - \ln 2$$

and thus:

$$\liminf_{T \to +\infty} \frac{\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]}{\ln T} \geqslant \frac{1-\alpha}{\mathrm{KL}(\nu_a, \nu'_a)}$$

We let $\alpha \downarrow 0$:
$$\liminf_{T \to +\infty} \frac{\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big]}{\ln T} \geqslant \frac{1}{\mathrm{KL}(\nu_a, \nu'_a)}$$

And then take the supremum of the right-hand side over $\nu'_a \in \mathcal{D}$ with $E(\nu'_a) > \mu^*$ to conclude.

<u>Other applications of the fundamental inequality</u> (see Garivier, Ménard, Stoltz, 2017)

among many, it's a very versatile tool that has already been used by other colleagues!

Optimal order for 2nd-order term for large T

* For "well-behaved" models $\mathcal{D}$ and strategies ensuring a $O(\ln T)$ bound on all $\mathbb{E}_{\underline{\nu}}[N_a(T)]$ when a is suboptimal for $\underline{\nu}$,

  we got a non-asymptotic lower bound of the form:
  $$\mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] \geqslant \frac{\ln T}{\mathcal{K}_{\inf}(\nu_a, \mu^*, \mathcal{D})} - \Omega(\ln \ln T)$$
  which matches an upper bound by Honda and Takemura, 2015, of the same form, for $\mathcal{D} = \mathcal{P}([0,1])$.

  Thus, at least for $\mathcal{D} = \mathcal{P}([0,1])$, the optimal order of magnitude of the second-order term is $-\ln \ln T$.

Linear bound on the regret for small T

* For all models $\mathcal{D}$, for all "reasonable" strategies (i.e. "smarter than the uniform strategy on $\mathcal{D}$"),

  for all bandit problems $\underline{\nu}$, for all arms a,
  $$\forall\, T \leqslant \frac{1}{8\, \mathcal{K}_{\inf}(\nu_a, \mu^*, \mathcal{D})}, \qquad \mathbb{E}_{\underline{\nu}}\big[N_a(T)\big] \geqslant \frac{T}{2K}$$

  i.e. uniform exploration of the arms at least half of the time!

Part 2:

Extracted from my M2 lecture notes

The Kullback-Leibler divergence: definition and basic properties.

Definition (intrinsic): Let $\mathbb{P}, \mathbb{Q}$ be two probability measures over $(\Omega, \mathcal{F})$

$$\mathrm{KL}(\mathbb{P},\mathbb{Q}) = \begin{cases} +\infty & \text{if } \mathbb{P} \text{ is not absolutely continuous wrt } \mathbb{Q} \\ \displaystyle\int_\Omega \left(\frac{d\mathbb{P}}{d\mathbb{Q}} \ln \frac{d\mathbb{P}}{d\mathbb{Q}} \right) d\mathbb{Q} = \int_\Omega \left(\ln \frac{d\mathbb{P}}{d\mathbb{Q}} \right) d\mathbb{P} & \text{if } \mathbb{P} \ll \mathbb{Q} \end{cases}$$

Basic facts:

- Existence of the defining integral when $\mathbb{P} \ll \mathbb{Q}$: because $\varphi : x \mapsto x \ln x$ is bounded from below on $[0,+\infty)$

- $\mathrm{KL}(\mathbb{P},\mathbb{Q}) \geq 0$ and $\mathrm{KL}(\mathbb{P},\mathbb{Q})$ if and only if $\mathbb{P} = \mathbb{Q}$:

It suffices to consider the case $\mathbb{P} \ll \mathbb{Q}$:

because φ is strictly convex, Jensen's inequality indicates that

$$\mathrm{KL}(\mathbb{P},\mathbb{Q}) = \int_\Omega \varphi\!\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q} \geq \varphi\!\left(\underbrace{\int_\Omega \frac{d\mathbb{P}}{d\mathbb{Q}} \, d\mathbb{Q}}_{=1} \right) = 0,$$

with equality if and only if $\frac{d\mathbb{P}}{d\mathbb{Q}}$ is $\mathbb{Q}$-as constant, ie, $\mathbb{P} = \mathbb{Q}$

Exercise 1: A useful rewriting. Prove the following result:

Assume $\mathbb{P} \ll \mathbb{Q}$ and let ν be any probability measure over $(\Omega, \mathcal{F})$ such that $\mathbb{P} \ll \nu$ and $\mathbb{Q} \ll \nu$. Denote $f = \frac{d\mathbb{P}}{d\nu}$ and $g = \frac{d\mathbb{Q}}{d\nu}$.

Then:

$$\mathrm{KL}(\mathbb{P},\mathbb{Q}) = \int_\Omega \frac{f}{g} \ln\!\left(\frac{f}{g}\right) g \, d\nu$$

$$= \int_\Omega \ln\!\left(\frac{f}{g}\right) f \, d\nu.$$

Beware: with the usual measure theoretic conventions, if $x \neq 0$ and $y = 0$, then $x \neq y \frac{x}{y}$

↳ you therefore need to proceed with care!

Lemma (contraction of entropy; also known as data-processing inequality):

Let $\mathbb{P}, \mathbb{Q}$ be two probability measures over $(\Omega, \mathcal{F})$

Let $X : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$ be any random variable

Denote by $\mathbb{P}^X$ and $\mathbb{Q}^X$ the laws of X under $\mathbb{P}$ and $\mathbb{Q}$.

Then:

$$\mathrm{KL}(\mathbb{P}^X, \mathbb{Q}^X) \leqslant \mathrm{KL}(\mathbb{P}, \mathbb{Q}).$$

Proof: We may assume that $\mathbb{P} \ll \mathbb{Q}$, otherwise $\mathrm{KL}(\mathbb{P},\mathbb{Q}) = +\infty$ and the inequality is true. We show that we then have

$$\mathbb{P}^X \ll \mathbb{Q}^X, \quad \text{with} \quad \frac{d\mathbb{P}^X}{d\mathbb{Q}^X} = \mathbb{E}_{\mathbb{Q}}\left[\frac{d\mathbb{P}}{d\mathbb{Q}} \,\Big|\, X = \cdot\,\right] \overset{\text{not.}}{=} \gamma$$

ie, $\gamma(X) = \mathbb{E}_{\mathbb{Q}}\left[\frac{d\mathbb{P}}{d\mathbb{Q}} \,\Big|\, X\right]$.

Indeed, for all $B \in \mathcal{F}'$:

$$\mathbb{P}^X(B) = \mathbb{P}\{X \in B\} = \int_\Omega \mathbb{1}_B(X) \frac{d\mathbb{P}}{d\mathbb{Q}}\, d\mathbb{Q} \overset{\text{tower rule}}{=} \int_\Omega \mathbb{1}_B(X)\, \mathbb{E}_{\mathbb{Q}}\left[\frac{d\mathbb{P}}{d\mathbb{Q}} \,\Big|\, X\right] d\mathbb{Q}$$

$$\overset{\text{not.}}{=} \int_\Omega \mathbb{1}_B(X)\, \gamma(X)\, d\mathbb{Q} \underset{\text{by definition of } \mathbb{Q}^X}{=} \int_{\Omega'} \mathbb{1}_B\, \gamma\, d\mathbb{Q}^X.$$

Therefore,

$$\mathrm{KL}(\mathbb{P}^X, \mathbb{Q}^X) = \int_{\Omega'} \gamma \ln \gamma \, d\mathbb{Q}^X = \int_\Omega \gamma(X) \ln \gamma(X)\, d\mathbb{Q}$$

definition of γ

$$= \int_\Omega \left(\mathbb{E}_{\mathbb{Q}}\left[\frac{d\mathbb{P}}{d\mathbb{Q}} \,\Big|\, X\right] \ln \mathbb{E}_{\mathbb{Q}}\left[\frac{d\mathbb{P}}{d\mathbb{Q}} \,\Big|\, X\right]\right) d\mathbb{Q}$$

conditional version of Jensen's inequality

$$\leqslant \int_\Omega \mathbb{E}_{\mathbb{Q}}\left[\frac{d\mathbb{P}}{d\mathbb{Q}} \ln \frac{d\mathbb{P}}{d\mathbb{Q}} \,\Big|\, X\right] d\mathbb{Q}$$

tower rule

$$= \int_\Omega \left(\frac{d\mathbb{P}}{d\mathbb{Q}} \ln \frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q} = \mathrm{KL}(\mathbb{P}, \mathbb{Q})$$

References : • The proof above is due to Ali and Silvey (1966), but it's far from being well-known!

• Typical proofs in the more recent literature:
  - either focus on the discrete case (Cover and Thomas, 2006)
  - or use the duality / variational formula for the KL (Massart, 2007 ; Boucheron, Lugosi, Massart, 2013)

• The joint convexity of KL, which we discuss below, is typically proved in a tedious way, relying on the rewriting of Exercise 1 and the joint convexity of $(x,y) \in [0,+\infty)^2 \longmapsto \left(\frac{x}{y} \ln \frac{x}{y}\right) y$

We may see it instead as a consequence of the data-processing inequality:

Corollary (joint convexity of KL): For all probability distributions $\mathbb{P}_1, \mathbb{P}_2$ and $\mathbb{Q}_1, \mathbb{Q}_2$ over the same measurable space $(\Omega, \mathcal{F})$, and all $\alpha \in (0,1)$,

$$\mathrm{KL}\big((1-\alpha)\mathbb{P}_1 + \alpha\mathbb{P}_2,\ (1-\alpha)\mathbb{Q}_1 + \alpha\mathbb{Q}_2\big) \leqslant (1-\alpha)\,\mathrm{KL}(\mathbb{P}_1, \mathbb{Q}_1) + \alpha\,\mathrm{KL}(\mathbb{P}_2, \mathbb{Q}_2)$$

Proof: We augment $(\Omega, \mathcal{F})$ into $(\Omega \times \{1,2\}, \mathcal{F}')$ where

$$\mathcal{F}' = \mathcal{F} \otimes \big\{ \emptyset, \{1\}, \{2\}, \{1,2\} \big\}$$

We define the random pair (X, J) by the projections

$$X : (\omega, j) \mapsto \omega \qquad \text{and} \qquad J : (\omega, j) \longmapsto j$$

Let $\mathbb{P}$ be a probability measure on $(\Omega \times \{1,2\}, \mathcal{F}')$ such that

$$\begin{cases} J \sim 1 + \mathrm{Ber}(\alpha) \\ X \,|\, J=j \sim \mathbb{P}_j \end{cases}$$

(and a similar definition for $\mathbb{Q}$ based on $\mathbb{Q}_1, \mathbb{Q}_2$)

that is, $\forall j \in \{1,2\}$ $\forall A \in \mathcal{F}$ $\mathbb{P}(A \times \{j\}) = \big((1-\alpha)\mathbb{1}_{\{j=1\}} + \alpha\,\mathbb{1}_{\{j=2\}}\big)\,\mathbb{P}_j(A)$

Now, $\mathbb{P}^X = (1-\alpha)\,\mathbb{P}_1 + \alpha \mathbb{P}_2$

$\mathbb{Q}^X = (1-\alpha)\,\mathbb{Q}_1 + \alpha \mathbb{Q}_2$

and (as we prove below) $\mathrm{KL}(\mathbb{P},\mathbb{Q}) = (1-\alpha)\,\mathrm{KL}(\mathbb{P}_1,\mathbb{Q}_1) + \alpha\,\mathrm{KL}(\mathbb{P}_2,\mathbb{Q}_2)$

so that the result follows from the data-processing inequality.

Indeed: we may assume with no loss of generality, given $\alpha \in (0,1)$, that $\mathbb{P}_1 \ll \mathbb{Q}_1$ and $\mathbb{P}_2 \ll \mathbb{Q}_2$, so that $\mathbb{P} \ll \mathbb{Q}$ with

$$\frac{d\mathbb{P}}{d\mathbb{Q}}(\omega,j) = \mathbb{1}_{\{j=1\}}\,\frac{d\mathbb{P}_1}{d\mathbb{Q}_1}(\omega) + \mathbb{1}_{\{j=2\}}\,\frac{d\mathbb{P}_2}{d\mathbb{Q}_2}(\omega)$$

This entails that (by Tonelli's theorem)

$$\mathrm{KL}(\mathbb{P},\mathbb{Q}) = \int_{\Omega\times\{1,2\}} \left(\frac{d\mathbb{P}}{d\mathbb{Q}}(\omega,j)\,\ln\frac{d\mathbb{P}}{d\mathbb{Q}}(\omega,j)\right) d\mathbb{Q}(\omega,j)$$

we integrate first over the j

$$= (1-\alpha)\int_{\Omega}\left(\frac{d\mathbb{P}}{d\mathbb{Q}}(\omega,1)\,\ln\frac{d\mathbb{P}}{d\mathbb{Q}}(\omega,1)\right) d\mathbb{Q}(\omega,1) + \alpha\int_{\Omega}\left(\frac{d\mathbb{P}}{d\mathbb{Q}}(\omega,2)\,\ln\frac{d\mathbb{P}}{d\mathbb{Q}}(\omega,2)\right) d\mathbb{Q}(\omega,2)$$

$$= (1-\alpha)\underbrace{\int_{\Omega}\left(\frac{d\mathbb{P}_1}{d\mathbb{Q}_1}(\omega)\,\ln\frac{d\mathbb{P}_1}{d\mathbb{Q}_1}(\omega)\right) d\mathbb{Q}_1(\omega)}_{=\mathrm{KL}(\mathbb{P}_1,\mathbb{Q}_1)} + \alpha\,\mathrm{KL}(\mathbb{P}_2,\mathbb{Q}_2)$$

KL for product measures. ($\leftrightarrow$ The independent case)

Proposition: Let $(\Omega,\mathcal{F})$ and $(\Omega',\mathcal{F}')$ be two measurable spaces, let $\mathbb{P},\mathbb{Q}$ be two probability measures over $(\Omega,\mathcal{F})$, $\mathbb{P}',\mathbb{Q}'$ over $(\Omega',\mathcal{F}')$ and denote by $\mathbb{P}\otimes\mathbb{P}'$ and $\mathbb{Q}\otimes\mathbb{Q}'$ the product distributions over $(\Omega\times\Omega', \mathcal{F}\otimes\mathcal{F}')$. Then:

$$\mathrm{KL}(\mathbb{P}\otimes\mathbb{P}', \mathbb{Q}\otimes\mathbb{Q}') = \mathrm{KL}(\mathbb{P},\mathbb{Q}) + \mathrm{KL}(\mathbb{P}',\mathbb{Q}')$$

Proof: It suffices to consider the case $\mathbb{P} \ll \mathbb{P}'$ and $\mathbb{Q} \ll \mathbb{Q}'$.

Then $\mathbb{P} \otimes \mathbb{P}' \ll \mathbb{Q} \otimes \mathbb{Q}'$ with

$$\frac{d(\mathbb{P} \otimes \mathbb{P}')}{d(\mathbb{Q} \otimes \mathbb{Q}')} = \frac{d\mathbb{P}}{d\mathbb{Q}} \, \frac{d\mathbb{P}'}{d\mathbb{Q}'}$$

(this is a fundamental result in measure theory and one of the best characterizations of independence!).

Therefore,

$$\mathrm{KL}(\mathbb{P} \otimes \mathbb{P}', \mathbb{Q} \otimes \mathbb{Q}') = \int_{\Omega \times \Omega'} \left(\frac{d\mathbb{P}}{d\mathbb{Q}} \frac{d\mathbb{P}'}{d\mathbb{Q}'} \ln\left(\frac{d\mathbb{P}}{d\mathbb{Q}} \frac{d\mathbb{P}'}{d\mathbb{Q}'} \right)\right) d(\mathbb{Q} \otimes \mathbb{Q}')$$

by Fubini-Tonelli (cf. $x \mapsto x \ln x$ is bounded from below)

$$= \underbrace{\int_{\Omega} \left(\underbrace{\int_{\Omega} \left(\frac{d\mathbb{P}}{d\mathbb{Q}} \ln \frac{d\mathbb{P}}{d\mathbb{Q}} \right) d\mathbb{Q}}_{= \mathrm{KL}(\mathbb{P},\mathbb{Q})} \right) \underbrace{\frac{d\mathbb{P}'}{d\mathbb{Q}'} d\mathbb{Q}'}_{d\mathbb{P}'}}_{= \mathrm{KL}(\mathbb{P},\mathbb{Q})} + \underbrace{\text{similar term with } \ln \frac{d\mathbb{P}'}{d\mathbb{Q}'}}_{= \mathrm{KL}(\mathbb{P}',\mathbb{Q}')}$$

Consequence (Garivier, Ménard, Stoltz, 2016):

Data-processing inequality with expectations of random variables

Corollary: Let $\mathbb{P}, \mathbb{Q}$ be two probability measures over $(\Omega, \mathcal{F})$

Let $X : (\Omega, \mathcal{F}) \to ([0,1], \mathcal{B}([0,1]))$ be any $[0,1]$-valued random variable

Then, denoting by $\mathbb{E}_{\mathbb{P}}[X]$ and $\mathbb{E}_{\mathbb{Q}}[X]$ the respective expectations of X under $\mathbb{P}$ and $\mathbb{Q}$, we have:

$$\mathrm{KL}\big(\mathrm{Ber}(\mathbb{E}_{\mathbb{P}}[X]), \mathrm{Ber}(\mathbb{E}_{\mathbb{Q}}[X])\big) \leqslant \mathrm{KL}(\mathbb{P}, \mathbb{Q})$$

Proof: We denote by $\mathfrak{m}$ the Lebesgue measure over $[0,1]$ and augment the underlying measurable space into $(\Omega \times [0,1], \mathcal{F} \otimes \mathcal{B}([0,1]))$, over which we consider the product-distributions $\mathbb{P} \otimes \mathfrak{m}$ and $\mathbb{Q} \otimes \mathfrak{m}$.

For any event $E \in \mathcal{F} \otimes \mathcal{B}([0,1])$, we have, by the data-processing inequality:

$$\mathrm{KL}\Big(\underbrace{(\mathbb{P}\otimes m)^{\mathbb{1}_E}}_{\mathrm{Ber}(\mathbb{P}\otimes m(E))},\ \underbrace{(\mathbb{Q}\otimes m)^{\mathbb{1}_E}}_{\mathrm{Ber}(\mathbb{Q}\otimes m(E))} \Big) \leqslant \mathrm{KL}(\mathbb{P}\otimes m, \mathbb{Q}\otimes m)$$

$$= \mathrm{KL}(\mathbb{P},\mathbb{Q}) + \mathrm{KL}(m,m)$$

↑ cf. product distributions

$$= \mathrm{KL}(\mathbb{P},\mathbb{Q})$$

Thus : $\quad \mathrm{KL}\big(\mathrm{Ber}(\mathbb{P}\otimes m(E)),\ \mathrm{Ber}(\mathbb{Q}\otimes m(E))\big) \leqslant \mathrm{KL}(\mathbb{P},\mathbb{Q})$

The proof is concluded by picking $E \in \mathcal{F}\otimes\mathcal{B}([0,1])$ such that

$$\mathbb{P}\otimes m(E) = \mathbb{E}_{\mathbb{P}}[X] \qquad \text{and} \qquad \mathbb{Q}\otimes m(E) = \mathbb{E}_{\mathbb{Q}}[X]$$

Namely, $\quad E = \{ (\omega, x) \in \Omega\times[0,1] : \ x \leqslant X(\omega) \}$

By Tonelli's theorem :

$$\mathbb{P}\otimes m(E) = \int_\Omega \left(\int_{[0,1]} \mathbb{1}_{\{x \leqslant X(\omega)\}}\, dm(x) \right) d\mathbb{P}(\omega)$$

$$= \int_\Omega X(\omega)\, d\mathbb{P}(\omega) = \mathbb{E}_{\mathbb{P}}[X]$$

and a similar equality for $\mathbb{Q}\otimes m(E)$.

<u>The chain rule</u> — A generalization of the decomposition of the KL between product distributions.

We will need it in a special case only, when the joint distributions follow from one of the marginal distributions via a stochastic kernel.

<u>Definition</u>: Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be two measurable spaces; we denote by $\mathcal{P}(\Omega', \mathcal{F}')$ the set of probability measures over $(\Omega', \mathcal{F}')$.

A stochastic kernel K is a mapping

$$(\Omega, \mathcal{F}) \longrightarrow \mathcal{P}(\Omega', \mathcal{F}')$$
$$\omega \longmapsto K(\omega, \cdot)$$

such that $\forall B \in \mathcal{F}'$ $\quad \omega \mapsto K(\omega, B)$ is $\mathcal{F}$-measurable.

Now, consider two such kernels K and L, and two probability measures $\mathbb{P}$ and $\mathbb{Q}$ over $(\Omega, \mathcal{F})$. Then $K\mathbb{P}$ and $L\mathbb{Q}$ defined below are probability measures over $(\Omega\times\Omega', \mathcal{F}\otimes\mathcal{F}')$:

$\forall A \in \mathcal{F}, \quad \forall B \in \mathcal{F}', \qquad K\mathbb{P}(A \times B) = \int_\Omega \mathbb{1}_A(\omega)\, K(\omega, B)\, d\mathbb{P}(\omega)$

$$L\mathbb{Q}(A \times B) = \int_\Omega \mathbb{1}_A(\omega)\, L(\omega, B)\, d\mathbb{Q}(\omega)$$

__Theorem__ (chain rule for KL):

As soon as $(*)$ $K(\omega, \cdot) \ll L(\omega, \cdot)$ for $\mathbb{P}$-almost all $\omega \in \Omega$,

with $(**)$ $g : (\omega, \omega') \longmapsto \dfrac{dK(\omega,\cdot)}{dL(\omega,\cdot)}(\omega')$ being $\mathcal{F} \otimes \mathcal{F}'$-measurable,

Then

$$\mathrm{KL}(K\mathbb{P}, L\mathbb{Q}) = \mathrm{KL}(\mathbb{P}, \mathbb{Q}) + \int_\Omega \mathrm{KL}\big(K(\omega,\cdot), L(\omega,\cdot)\big)\, d\mathbb{P}(\omega)$$

where $\omega \mapsto \mathrm{KL}\big(K(\omega,\cdot), L(\omega,\cdot)\big)$ is indeed $\mathcal{F}$-measurable, so that the integral in the right-hand side is well defined.

__Remarks:__ Assumptions $(*)$ and $(**)$ will be fulfilled when applying this equality to our stochastic bandit setting.

They are actually unnecessary as proved for me by Nath Balk (ENS student following my M2 classes this year), at least in the most common cases: see details after the proof!

__Proof:__ • Since $g \ln g$ is lower bounded on $[0, +\infty)$, and in view of the measurability assumption $(**)$, Tonelli's theorem wrt $L\mathbb{Q}$ ensures that

$$\omega \longmapsto \int_{\Omega'} \big(g(\omega,\omega') \ln g(\omega,\omega')\big)\, L(\omega, d\omega') = \mathrm{KL}\big(K(\omega,\cdot), L(\omega,\cdot)\big)$$

is $\mathcal{F}$-measurable.

• Given $(*)$, we have $K\mathbb{P} \ll L\mathbb{Q}$ if and only if $\mathbb{P} \ll \mathbb{Q}$; we thus assume with no loss of generality that $K\mathbb{P} \ll L\mathbb{Q}$ and $\mathbb{P} \ll \mathbb{Q}$ (otherwise, both sides of the putative equality equal $+\infty$).

• We write $\dfrac{d\mathbb{P}}{d\mathbb{Q}} = f$, we then have

$$\frac{d\, K\mathbb{P}}{d\, L\mathbb{Q}}(\omega,\omega') = f(\omega)\; g(\omega,\omega')$$

as can be seen by going back to the definition of $L\mathbb{Q}$

And
$$KL(K\mathbb{P}, L\mathbb{Q}) = \int_{\Omega\times\Omega'} \Big(f(\omega)\, g(\omega,\omega') \ln\big(f(\omega) g(\omega,\omega')\big)\Big)\; \underbrace{d\, L\mathbb{Q}(\omega,\omega')}_{L(\omega,\, d\omega')\, d\mathbb{Q}(\omega)}$$

by Tonelli's theorem

$$= \int_{\Omega} f(\omega) \ln f(\omega) \Big(\underbrace{\int_{\Omega'} g(\omega,\omega')\, L(\omega, d\omega')}_{=K(\omega,\Omega')=1} \Big)\, d\mathbb{Q}(\omega)$$

$$+ \int_{\Omega} \Big(\underbrace{\int_{\Omega'} \big(g(\omega,\omega') \ln g(\omega,\omega')\big)\, L(\omega, d\omega')}_{KL(K(\omega,\cdot),\, L(\omega,\cdot))} \Big)\; \underbrace{f(\omega)\, d\mathbb{Q}(\omega)}_{d\mathbb{P}(\omega)}$$

$$= \underbrace{\int_{\Omega} f \ln f \; d\mathbb{Q}}_{=KL(\mathbb{P},\mathbb{Q})} + \int_{\Omega} KL\big(K(\omega,\cdot),\, L(\omega,\cdot)\big)\; d\mathbb{P}(\omega)$$

as announced.

Remark: We can prove that $K\mathbb{P} \ll L\mathbb{Q} \iff \begin{cases} \mathbb{P} \ll \mathbb{Q} \\ K(\omega,\cdot) \ll L(\omega,\cdot) \text{ for } \mathbb{P}\text{-almost all } \omega\in\Omega \end{cases}$

holds as soon as Ω' is a topological space with a countable base and $\mathcal{F}'$ is the Borel σ-algebra.

This is, for instance, the case when Ω' is a separable metric space.

Consequences of this equivalence: The putative equality is $+\infty = +\infty$ if $K\mathbb{P} \not\ll L\mathbb{Q}$.
We may thus assume with no loss of generality that $K\mathbb{P} \ll L\mathbb{Q}$, with density f
$\mathbb{P} \ll \mathbb{Q}$, with density h
and $K(\omega,\cdot) \ll L(\omega,\cdot)$ for $\mathbb{P}$-almost all $\omega\in\Omega$.

This is $(*)$. As for $(**)$, we may then prove that

for $\mathbb{P}$-almost all $\omega\in\Omega$, $\displaystyle\frac{dK(\omega,\cdot)}{dL(\omega,\cdot)} = \frac{f(\omega,\cdot)}{h(\omega)}$

where f/h is $\mathcal{F}\otimes\mathcal{F}'$-measurable.

Part 3: Detailed proofs of the lower bounds.

Reminder: • K arms $k = 1, 2, \ldots K$ each associated with a probability distribution $\nu_1, \ldots \nu_K$ lying in some bandit model $\mathcal{D}$

with expectations $\mu_1, \ldots \mu_K$

Optimal arms: arms a s.t. $\mu_a = \mu^* = \max_{k=1..K} \mu_k$
Suboptimal arms: $\mu_a < \mu^*$; we denote $\Delta_a = \mu^* - \mu_a$ the gap

• Randomized strategies $(\Psi_t)_{t \geq 0}$, where

$$\Psi_t : H_t = (U_0, Y_1, U_1, \ldots Y_t, U_t) \longmapsto \Psi_t(H_t) = I_{t+1}$$

• Laws: $Y_t \mid I_t \sim \nu_{I_t} \quad \forall t \geq 1$ and $U_0, U_1, \ldots$ iid $\sim \mathcal{U}_{[0,1]}$

• Regret: $R_T = T\mu^* - \mathbb{E}_{\underline{\nu}}\left[\sum_{t=1}^{T} Y_t\right] \underset{\text{tower rule}}{=} \sum_{a=1}^{K} \Delta_a \, \mathbb{E}_{\underline{\nu}}[N_a(T)]$

where $\mathbb{E}_{\underline{\nu}}$ denotes the expectation when the underlying bandit problem is $\underline{\nu} = (\nu_a)_{a=1..K}$

↳ To control the regret (upper or lower bound it), it is thus necessary and sufficient to control the $\mathbb{E}_{\underline{\nu}}[N_a(T)]$.

and where $N_a(T) = \sum_{t=1}^{T} \mathbb{1}_{\{I_t = a\}}$

Lemma [Fundamental inequality for stochastic bandits]:

For all bandit models $\mathcal{D}$, for all strategies Ψ,

For all bandit problems $\underline{\nu} = (\nu_a)_{a=1..K}$ and $\underline{\nu}' = (\nu'_a)_{a=1..K}$ in $\mathcal{D}$ with $\nu_a \ll \nu'_a \; \forall a$,

For all random variables Z taking values in $[0,1]$ and that are $\sigma(H_T)$-measurable,

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

$$\geq \mathrm{KL}\left(\mathrm{Ber}(\mathbb{E}_{\underline{\nu}}[Z]), \mathrm{Ber}(\mathbb{E}_{\underline{\nu}'}[Z])\right)$$

where $\mathbb{P}_{\underline{\nu}}^{H_T}$ and $\mathbb{P}_{\underline{\nu}'}^{H_T}$ denote the laws of H_T when the strategy is Ψ and when the underlying bandit problems are respectively $\underline{\nu}$ and $\underline{\nu}'$.

Proof: • The inequality $\geqslant$ is a direct application of the data-processing inequality with expectations, see the previous lecture for its statement.

• For the equality: We will explain how $\mathbb{P}_{\nu}^{H_T}$ is constructed.

With no loss of generality, we can consider that

- the underlying probability space is $\Omega = [0,1] \times (\mathbb{R} \times [0,1])^T$
- H_T is the identity over Ω, i.e., that the $U_0, Y_1, U_1, \ldots Y_T, U_T$ are the projections on each component,
- $\mathbb{P}_{\nu}$ is given by

$$\forall B \in \mathcal{B}([0,1]), \qquad \mathbb{P}_{\nu}(U_0 \in B) = \eta(B)$$

$$\forall t \in \{0, \ldots T-1\}, \quad \forall B' \in \mathcal{B}(\mathbb{R}), \ \forall B \in \mathcal{B}([0,1]), \qquad \mathbb{P}_{\nu}\big(Y_{t+1} \in B' \text{ and } U_{t+1} \in B \mid H_t\big) = \nu_{\psi_t(H_t)}(B)\,\eta(B')$$

where $\mathcal{B}(S)$ is the Borel-σ-algebra of a set $S \subseteq \mathbb{R}$
η is the Lebesgue measure over $[0,1]$

In particular: $\mathbb{P}_{\nu}^{H_t}$ refers to the first $1+2t$ marginals of $\mathbb{P}_{\nu}^{H_T} = \mathbb{P}_{\nu}$.

A similar construction can be done for the bandit problem ν'.

Now, the equality $\mathbb{P}_{\nu}\big(Y_{t+1} \in B' \text{ and } U_{t+1} \in B \mid H_t\big) = \nu_{\psi_t(H_t)}(B)\ \eta(B)$ indicates that $\mathbb{P}_{\nu}^{H_{t+1}} = K_t\, \mathbb{P}_{\nu}^{H_T}$ for the regular transition kernel

$$K_t(h,\cdot) = \nu_{\psi_t(h)} \otimes \eta$$

regularity is: for $E \in \mathcal{B}(\mathbb{R}) \otimes \mathcal{B}([0,1])$, $h \mapsto K_t(h,E)$ is measurable

Similarly, $\mathbb{P}_{\nu'}^{H_{t+1}} = K'_t\, \mathbb{P}_{\nu'}^{H_t}$ for $K'_t(h,\cdot) = \nu'_{\psi_t(h)} \otimes \eta$

Let us check the assumptions of our chain rule:

(*) $\quad \forall h, \quad K_t(h,\cdot) \ll K'_t(h,\cdot) \quad$ as $\forall a,\ \nu_a \ll \nu'_a$ by assumption

(**) $\quad (h,(y,u)) \longmapsto \dfrac{dK_t(h,\cdot)}{dK'_t(h,\cdot)}(y,u) = \displaystyle\sum_{a=1}^{K} \mathbb{1}_{\{\psi_t(h)=a\}} \frac{d\nu_a}{d\nu'_a}(y)$
is indeed bi-measurable.

Therefore, for $t \in \{0, \ldots T-1\}$,

$$\mathrm{KL}\big(\mathbb{P}_{\nu}^{H_{t+1}}, \mathbb{P}_{\nu'}^{H_{t+1}}\big) = \mathrm{KL}\big(\mathbb{P}_{\nu}^{H_t}, \mathbb{P}_{\nu'}^{H_t}\big) + \int \mathrm{KL}\big(\nu_{\psi_t(h)} \cancel{\otimes \eta}, \nu'_{\psi_t(h)} \cancel{\otimes \eta}\big)\, d\mathbb{P}_{\nu}^{H_t}(h)$$

$$= \mathrm{KL}\big(\mathbb{P}_{\nu}^{H_t}, \mathbb{P}_{\nu'}^{H_t}\big) + \sum_{a=1}^{K} \mathrm{KL}(\nu_a, \nu'_a)\ \mathbb{P}_{\nu}^{H_t}\{\psi_t(h) = a\}$$

Now, $I_{t+1} = \psi_t(H_t)$ so that

$$\mathbb{P}_{\nu}^{H_t}\{\psi_t(h) = a\} = \mathbb{P}_{\nu}\{\psi_t(H_t) = a\} = \mathbb{P}_{\nu}\{I_{t+1} = a\}$$
$$= \mathbb{E}_{\nu}\left[\mathbb{1}_{\{I_{t+1} = a\}}\right]$$

Summing up:

- $KL\left(\mathbb{P}_{\nu}^{H_0}, \mathbb{P}_{\nu'}^{H_0}\right) = KL\left(\mathbb{P}_{\nu}^{U_0}, \mathbb{P}_{\nu'}^{U_0}\right) = KL(\eta, \eta) = 0$
- $\forall t \in \{0, \dots T-1\}$, $KL\left(\mathbb{P}_{\nu}^{H_{t+1}}, \mathbb{P}_{\nu'}^{H_{t+1}}\right) = KL\left(\mathbb{P}_{\nu}^{H_t}, \mathbb{P}_{\nu'}^{H_t}\right) + \sum_{a=1}^{K} KL(\nu_a, \nu'_a)\, \mathbb{E}_{\nu}\left[\mathbb{1}_{\{I_{t+1} = a\}}\right]$

so that the stated result follows by induction.

<u>Back to the theorem giving the distribution-dependent lower bound:</u>

<u>Theorem 1:</u> For all bandit models $\mathcal{D}$,
For all (possibly randomized) strategies ψ uniformly fast convergent on $\mathcal{D}$,
For all bandit problems $\underline{\nu} = (\nu_a)_{a \in \{1..K\}}$ in $\mathcal{D}$,
For all suboptimal arms a (ie, arms a with $\Delta_a > 0$),

$$\liminf_{T \to +\infty} \frac{\mathbb{E}_{\underline{\nu}}[N_a(T)]}{\ln T} \geqslant \frac{1}{K_{\inf}(\nu_a, \mu^*, \mathcal{D})}$$

where $K_{\inf}(\nu_a, \mu^*, \mathcal{D}) = \inf\left\{ KL(\nu_a, \nu'_a) : \nu'_a \in \mathcal{D} \text{ with } E(\nu'_a) > \mu^* \right\}$
with the convention $\inf \emptyset = +\infty$.

<u>Definition</u> (recalled): A strategy ψ is uniformly fast convergent on a model $\mathcal{D}$ if
for all bandit problems $\underline{\nu} = (\nu_k)_{k=1..K}$ in $\mathcal{D}$,
for all suboptimal arms a,
for all $\alpha \in (0,1]$,

$$\mathbb{E}_{\underline{\nu}}[N_a(T)] = o(T^{\alpha}).$$

Proof: We have $K_{\inf}(\nu_a, \mu^*) = \inf \{ KL(\nu_a, \nu'_a) : \nu'_a \in \mathcal{D} \text{ and } E(\nu'_a) > \mu^* \}$

$= \inf \{ KL(\nu_a, \nu'_a) : \nu'_a \in \mathcal{D}, \ \nu_a \ll \nu'_a \text{ and } E(\nu'_a) > \mu^* \}$

(cf. convention: $\inf \emptyset = +\infty$ and the fact that $KL(\nu_a, \nu'_a) = +\infty$ when $\nu_a \not\ll \nu'_a$)

This is why we will

- Fix $\mathcal{D}$, ψ, ν and a s.t. $\Delta_a > 0$
- Fix an alternative model ν' of the form

$$\begin{cases} \nu'_k = \nu_k & \forall k \neq a \\ \nu'_a & \text{s.t. } \nu'_a \in \mathcal{D}, \ \nu_a \ll \nu'_a \text{ and } E(\nu'_a) > \mu^* \end{cases}$$

That is, ν and ν' only differ at a; a is the unique optimal arm in ν'

- Take $Z = N_a(T)/T$ which is indeed $[0,1]$-valued, $\sigma(H_T)$-measurable

Our fundamental inequality yields, since ν and ν' only differ at a:

$$\mathbb{E}_\nu[N_a(T)] \ KL(\nu_a, \nu'_a) \geq KL\left(Ber\left(\mathbb{E}_\nu[N_a(T)/T]\right), Ber\left(\mathbb{E}_{\nu'}\left[\frac{N_a(T)}{T}\right]\right)\right)$$

$$\geq -\ln 2 + \left(1 - \mathbb{E}_\nu[N_a(T)/T]\right) \ln \frac{1}{1 - \mathbb{E}_{\nu'}[N_a(T)/T]}$$

indeed: $KL(Ber(p), Ber(q))$

$= p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}$

$= \underbrace{p \ln \frac{1}{q}}_{\geq 0} + (1-p) \ln \frac{1}{1-q} + \underbrace{(p \ln p + (1-p) \ln(1-p))}_{\geq -\ln 2 \text{ by a simple function study over } [0,1]}$

$\geq -\ln 2 + (1-p) \ln \frac{1}{1-q}$

for all $p, q \in (0,1)$ and even for all $p, q \in [0,1]$ (study the cases $q=0$ and $q=1$ separately)

Now, the considered strategy ψ is consistent and:

- in the problem ν, a is suboptimal: $\mathbb{E}_\nu[N_a(T)/T] \longrightarrow 0$

– in the problem $\underline{\nu}'$, all arms $k \neq a$ are suboptimal:

for all $\alpha \in (0,1]$, $\quad T - \mathbb{E}_{\underline{\nu}'}[N_a(T)] = \underbrace{\sum_{k \neq a} \mathbb{E}_{\underline{\nu}'}[N_k(T)]}_{= o(T^\alpha)}$

↳ in particular, for T large enough,

$$\frac{1}{1 - \mathbb{E}_{\underline{\nu}'}[N_a(T)/T]} = \frac{T}{T - \mathbb{E}_{\underline{\nu}'}[N_a(T)]} \geqslant \frac{T}{T^\alpha} = T^{1-\alpha}$$

Substituting back and dividing by $\ln T$: for all $\alpha \in (0,1]$, for T large enough:

$$\frac{\mathbb{E}_{\underline{\nu}}[N_a(T)] \quad KL(\nu_a, \nu'_a)}{\ln T} \geqslant -\frac{\ln 2}{\ln T} + \Big(1 - \underbrace{\mathbb{E}_{\underline{\nu}}\Big[\frac{N_a(T)}{T}\Big]}_{\to 0}\Big) \underbrace{\frac{\ln T^{1-\alpha}}{\ln T}}_{= 1-\alpha}$$

thus

$$\liminf_{T \to +\infty} \frac{\mathbb{E}_{\underline{\nu}}[N_a(T)] \quad KL(\nu_a, \nu'_a)}{\ln T} \geqslant 1 - \alpha$$

Letting $\alpha \to 0$, $\quad \displaystyle\liminf_{T \to +\infty} \frac{\mathbb{E}_{\underline{\nu}}[N_a(T)] \quad KL(\nu_a, \nu'_a)}{\ln T} \geqslant 1$

Whether $KL(\nu_a, \nu'_a) < +\infty$ or $= +\infty$, we thus get

$$\liminf_{T \to +\infty} \frac{\mathbb{E}_{\underline{\nu}}[N_a(T)]}{\ln T} \geqslant \frac{1}{KL(\nu_a, \nu'_a)}$$

The left-hand side is independent of $\nu'_a \in \mathcal{D}$ s.t. $\nu'_a \gg \nu_a$ and $E(\nu'_a) > \mu^*$, so that taking the supremum of the right-hand side over these ν'_a, we get the desired $1/K_{\inf}(\nu_a, \mu^*)$ lower bound.

Corollary [back to the regret!]:

For all bandit models $\mathcal{D}$,

For all randomized strategies ψ uniformly fast convergent on $\mathcal{D}$,

For all bandit problems $\underline{\nu} = (\nu_a)_{a=1..K}$ in $\mathcal{D}$,

$$\liminf_{T \to +\infty} \frac{R_T}{\ln T} \geqslant \sum_{a \,:\, \Delta_a > 0} \frac{\Delta_a}{K_{\inf}(\nu_a, \mu^*, \mathcal{D})}$$

Now, let's move to the distribution-free (minimax) lower bound:

Theorem 2: For all $K \geq 2$ and $T \geq K/5$

$$\inf_{\substack{\text{strategies}\\ \psi}} \sup_{\substack{\nu_1,\dots,\nu_K\\ \text{in } \mathcal{P}([0,1])}} R_T \geq \frac{1}{20}\sqrt{TK}.$$

and even: sup over $\nu_1 \dots \nu_K$ being Bernoulli distributions

Proof: • We consider the bandit problem $\underline{\nu}^{(0)} = \big(\mathrm{Ber}(1/2 - \varepsilon/2), \dots \mathrm{Ber}(1/2 - \varepsilon/2)\big)$

versus the bandit problems $\underline{\nu}^{(i)} = \big(\mathrm{Ber}(1/2 - \varepsilon/2), \dots \mathrm{Ber}(1/2 - \varepsilon/2), \underset{\uparrow \text{ i-th position}}{\mathrm{Ber}(1/2 + \varepsilon/2)}, \mathrm{Ber}(1/2 - \varepsilon/2), \dots \mathrm{Ber}(1/2 - \varepsilon/2)\big)$
for $i \in \{1..K\}$ and $\varepsilon \in (0, 1/2)$

For all strategies,

$$\sup_{\underline{\nu} \text{ in } \mathcal{P}([0,1])} R_T \geq \sup_{\varepsilon \in (0,1/2)} \max_{i \in \{1..K\}} R_T$$

$$= \sup_{\varepsilon \in (0,1/2)} \max_{i \in \{1..K\}} \sum_{k \neq i} \varepsilon \, \mathbb{E}_{\underline{\nu}^{(i)}}\big[N_k(T)\big]$$

ε: gap of arm $k \neq i$ in $\underline{\nu}^{(i)}$

$\mathbb{E}_{\underline{\nu}^{(i)}}[N_k(T)]$: expected number of times k is pulled under $\underline{\nu}^{(i)}$

And $\displaystyle\sum_{k \neq i} N_k(T) = T - N_i(T)$

Thus, $$\sup_{\underline{\nu} \text{ in } \mathcal{P}([0,1])} R_T \geq \sup_{\varepsilon \in (0,1/2)} \max_{i \in \{1..K\}} \varepsilon\Big(T - \mathbb{E}_{\underline{\nu}^{(i)}}\big[N_i(T)\big]\Big)$$

• Now, we use the fundamental inequality to upper bound one of the $\mathbb{E}_{\underline{\nu}^{(i)}}[N_i(T)]$

There exists $k \in \{1..K\}$ such that $\mathbb{E}_{\underline{\nu}^{(0)}}[N_k(T)] \leq T/K$. For this k:

$$\sum_{a=1}^{K} \mathbb{E}_{\underline{\nu}^{(0)}}\big[N_a(T)\big] \, \mathrm{KL}\big(\nu_a^{(0)}, \nu_a^{(k)}\big) \geq \mathrm{kl}\Big(\mathbb{E}_{\underline{\nu}^{(0)}}\big[N_k(T)/T\big], \mathbb{E}_{\underline{\nu}^{(k)}}\big[N_k(T)/T\big]\Big)$$

since $\nu^{(0)}$ and $\nu^{(k)}$ only differ at arm k:

$$= \mathbb{E}_{\underline{\nu}^{(0)}}\big[N_k(T)\big] \, \mathrm{KL}\big(\mathrm{Ber}(1/2 - \varepsilon/2), \mathrm{Ber}(1/2 + \varepsilon/2)\big)$$

$$\leq \underbrace{\frac{T}{K}}_{\text{by definition of } k} \mathrm{kl}\big(1/2 - \varepsilon/2, 1/2 + \varepsilon/2\big)$$

We lower bound this side by Pinsker's inequality for Bernoulli distributions:

$$\geq 2\left(\mathbb{E}_{\underline{\nu}^{(0)}}\left[\frac{N_k(T)}{T}\right] - \mathbb{E}_{\underline{\nu}^{(k)}}\left[\frac{N_k(T)}{T}\right]\right)^2$$

We proved so far:

$$2\left(\mathbb{E}_{\underline{\nu}^{(0)}}\left[\frac{N_k(T)}{T}\right] - \mathbb{E}_{\underline{\nu}^{(k)}}\left[\frac{N_k(T)}{T}\right]\right)^2 \leq \frac{T}{K} \mathrm{kl}\big(1/2 - \varepsilon/2, 1/2 + \varepsilon\big)$$

This entails in particular :

$$\mathbb{E}_{\underline{\nu}(k)}\left[\frac{N_k(T)}{T}\right] \leqslant \underbrace{\mathbb{E}_{\underline{\nu}(0)}\left[\frac{N_k(T)}{T}\right]}_{\leqslant 1/K \text{ by definition of } k} + \sqrt{\frac{T}{2K}\,\mathrm{kl}(1/2-\varepsilon/2,\ 1/2+\varepsilon/2)}$$

- Substituting in the regret lower bound, we get (with $i=k$) :

$$\sup_{\underline{\nu} \text{ in } \mathcal{D}([0,1])} R_T \geqslant \sup_{\varepsilon\in(0,1/2)} \varepsilon T\left(1-\mathbb{E}_{\underline{\nu}(k)}\left[\frac{N_k(T)}{T}\right]\right)$$

$$\geqslant \sup_{\varepsilon\in(0,1/2)} \varepsilon T\left(\underbrace{1-1/K}_{\geqslant 1/2} - \sqrt{\frac{T}{2K}\,\mathrm{kl}(1/2-\varepsilon/2,\ 1/2+\varepsilon/2)}\right)$$

$$\geqslant \sup_{\varepsilon\in(0,1/2)} \varepsilon T\left(1/2 - \sqrt{\frac{T}{2K}\,\mathrm{kl}(1/2-\varepsilon/2,\ 1/2+\varepsilon/2)}\right)$$

Now,

$$\mathrm{kl}(1/2-\varepsilon/2,\ 1/2+\varepsilon/2) = (\cancel{1/2}-\varepsilon/2)\ln\frac{1-\varepsilon}{1+\varepsilon} + (\cancel{1/2}+\varepsilon/2)\ln\frac{1+\varepsilon}{1-\varepsilon}$$

$$= \varepsilon \ln\frac{1+\varepsilon}{1-\varepsilon}$$

$$\leqslant 2.5\ \varepsilon^2 \text{ on } (0,1/2)$$

(study $\varepsilon \mapsto \varepsilon\ln\left(\frac{1+\varepsilon}{1-\varepsilon}\right)/\varepsilon^2$)

So that the regret lower bound becomes

$$\sup_{\underline{\nu} \text{ in } \mathcal{D}([0,1])} R_T \geqslant \sup_{\varepsilon\in(0,1/2)} \frac{T}{2}\left(\varepsilon - \varepsilon^2\sqrt{5T/K}\right)$$

we pick ε s.t. $1-2\varepsilon\sqrt{5T/K}=0$,

that is, $\varepsilon = \sqrt{K/T}\big/2\sqrt{5}$

which is indeed $< 1/2$ as soon as $T > K/5$.

We get a $\sqrt{TK}\left(1/2\left(1/2\sqrt{5} - \frac{1}{20}\sqrt{5}\right)\right) \geqslant \frac{1}{20}\sqrt{KT}$ bound.

Note : The original proof used $\mathrm{Ber}(1/2)$ and $\mathrm{Ber}(1/2+\varepsilon)$ but the approach with $\mathrm{Ber}(1/2-\varepsilon)$ and $\mathrm{Ber}(1/2+\varepsilon)$ leads to slightly simpler calculation.