Part 1: The Kullback-Leibler divergence, more properties

Recap of what we already saw on KL-divergences:

Definition: $KL(\mathbb{P},\mathbb{Q}) = \begin{cases} +\infty & \text{if } \mathbb{P} \not\ll \mathbb{Q} \\ \int_\Omega \left(\frac{d\mathbb{P}}{d\mathbb{Q}} \ln \frac{d\mathbb{P}}{d\mathbb{Q}} \right) d\mathbb{Q} & \text{if } \mathbb{P} \ll \mathbb{Q} \end{cases}$

First properties: $KL(\mathbb{P},\mathbb{Q}) \geqslant 0$ with $KL(\mathbb{P},\mathbb{Q}) = 0 \iff \mathbb{P} = \mathbb{Q}$

Data-processing inequality (with X r.v.): $KL(\mathbb{P}^X, \mathbb{Q}^X) \leq KL(\mathbb{P},\mathbb{Q})$

Joint convexity of KL: $KL\big((1-\alpha)\mathbb{P}_1 + \alpha\mathbb{P}_2, (1-\alpha)\mathbb{Q}_1 + \alpha\mathbb{Q}_2\big) \leqslant (1-\alpha)\, KL(\mathbb{P}_1,\mathbb{Q}_1) + \alpha\, KL(\mathbb{P}_2,\mathbb{Q}_2)$

First new result today:

KL for product measures ($\leftrightarrow$ the independent case, while the dependent case will be considered in a few pages)

Prop: Let $(\Omega,\mathcal{F})$ and $(\Omega',\mathcal{F}')$ be two measurable spaces, let $\mathbb{P},\mathbb{Q}$ be two probability measures over $(\Omega,\mathcal{F})$ and $\mathbb{P}',\mathbb{Q}'$ 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:

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

Proof: We have $\mathbb{P}\otimes\mathbb{P}' \ll \mathbb{Q}\otimes\mathbb{Q}' \hookrightarrow [\mathbb{P}\ll\mathbb{Q} \text{ and } \mathbb{P}'\ll\mathbb{Q}']$

So we can assume that all $\ll$ statements hold, and then

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

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

Therefore,

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

We use that if f, g are $\geq$ a constant, then $\int(f+g)d\mu = \int f d\mu + \int g d\mu$

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

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

here we apply Tonelli's theorem (again because $x\mapsto x\ln x$ is lower bounded)

Consequence (Gerber, Nénard, 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:

$$\mathbb{E}_{\mathbb{P}}[X]\ln\frac{\mathbb{E}_{\mathbb{P}}[X]}{\mathbb{E}_{\mathbb{Q}}[X]} + (1-\mathbb{E}_{\mathbb{P}}[X])\ln\frac{1-\mathbb{E}_{\mathbb{P}}[X]}{1-\mathbb{E}_{\mathbb{Q}}[X]} = KL\left(Ber(\mathbb{E}_{\mathbb{P}}[X]), Ber(\mathbb{E}_{\mathbb{Q}}[X])\right) \leq KL(\mathbb{P},\mathbb{Q})$$

Proof: We denote by η 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 $\mathbb{P}\otimes\eta$ and $\mathbb{Q}\otimes\eta$.

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

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

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

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

(if product distributions)

Thus: $KL\Big(\mathrm{Ber}(\mathbb{P}\ll m(E)), \mathrm{Ber}(\mathbb{Q}\ll m(E))\Big) \leq KL(\mathbb{P},\mathbb{Q})$

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

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

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

By Tonelli's theorem:

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

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

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

<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 (regular) K is a mapping $(\Omega,\mathcal{F}) \to \mathcal{P}(\Omega',\mathcal{F}')$, $\omega \mapsto K(\omega,\cdot)$

such that $\forall B\in\mathcal{F}'$ $\omega \mapsto K(\omega,B)$ 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 $(\Omega\times\Omega', \mathcal{F}\otimes\mathcal{F}')$, by some extension theorem: (Carathéodory?)

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

($\mathbb{1}_A(\omega)\, K(\omega,B)$ is indeed measurable)

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

An extension of Fubini(-Tonelli) theorem

Lemma: Let $\varphi : \Omega\times\Omega' \to \mathbb{R}$ be either ≥ 0 or $K\otimes P$-integrable. ($\mathcal{F}\otimes\mathcal{F}'$-measurable and lower bounded by $m\in\mathbb{R}$)

Then $\omega \mapsto \int_{\Omega'} \varphi(\omega,\omega')\, K(\omega, d\omega')$ is $\mathcal{F}$-measurable and

$$\int_{\Omega\times\Omega'} \varphi \, dK\otimes P = \int_\Omega \left(\int_{\Omega'} \varphi(\omega,\omega')\, K(\omega, d\omega') \right) dP(\omega)$$

Proof: (sketch) The result is true for $\varphi = \mathbb{1}_{A\times B}$ by definition of $K\otimes P$ (including measurability of $\omega \mapsto \int \varphi(\omega,\cdot)\, K(\omega, d\cdot)$ by regularity of K)

Extension to $\mathbb{1}_E$ for any $E \in \mathcal{F}\otimes\mathcal{F}'$ by monotone class theorem, using σ-algebra contained / monotone convergence (including the $\omega \mapsto \int_{\Omega'} \ldots$ measurability)

Extension to $\begin{cases} \varphi \geq 0 \\ \varphi \in L^1 \end{cases}$ by monotone convergence.

Question: Does anyone have a simpler argument?

Theorem [chain rule for KL]:

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

with (**) the existence of a version $g : (\omega,\omega') \mapsto \frac{dK(\omega,\cdot)}{dL(\omega,\cdot)}(\omega')$ being $\mathcal{F}\otimes\mathcal{F}'$-measurable,

Then

$$KL(KP, LQ) = KL(P,Q) + \int_\Omega KL(K(\omega,\cdot), L(\omega,\cdot))\, dP(\omega)$$

where $\omega \mapsto KL(K(\omega,\cdot), L(\omega,\cdot))$ is indeed $\mathcal{F}$-measurable and ≥ 0 so that the integral in the right-hand side is well defined.

Remark: see a remark stated in two pages for the (lack of) necessity of Assumptions (*) and (**).

Proof: * By bi-measurability of $g \ln g$, and since $g \ln g$ is lower bounded, (an immediate extension of) the previous lemma can be applied to get

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

is $\mathcal{F}$-measurable and ≥ 0, with:

(we will not use this, actually)

$$\int_{\Omega\times\Omega'} g \ln g \, d\,\mathbb{LP} = \int KL(K(\omega,\cdot), L(\omega,\cdot))\, d\mathbb{P}(\omega)$$

* We assume $\mathbb{P} \ll \mathbb{Q}$, let $f = \frac{d\mathbb{P}}{d\mathbb{Q}}$: what can we say about $(\omega,\omega') \mapsto f(\omega)\, g(\omega,\omega')$?

$$\int \mathbb{1}_{A\times B}(\omega,\omega')\, f(\omega)\, g(\omega,\omega')\, d\mathbb{LQ}(\omega,\omega')$$

$$= \text{(by extension of Tonelli)} \int_{\Omega} \left(\int_{\Omega'} \mathbb{1}_B(\omega')\, g(\omega,\omega')\, L(\omega, d\omega') \right) \mathbb{1}_A(\omega)\, f(\omega)\, d\mathbb{Q}(\omega)$$

$$\int_{\Omega'} \mathbb{1}_B(\omega')\, K(\omega, d\omega') = K(\omega, B)$$

given the definition of g

$$= \int \underbrace{\mathbb{1}_A(\omega)\, K(\omega,B)}_{\mathcal{F}\text{-measurable}}\, \underbrace{f(\omega)\, d\mathbb{Q}(\omega)}_{d\mathbb{P}(\omega)} = K\mathbb{P}(A\times B) \quad \text{by def. of } K\mathbb{P}$$

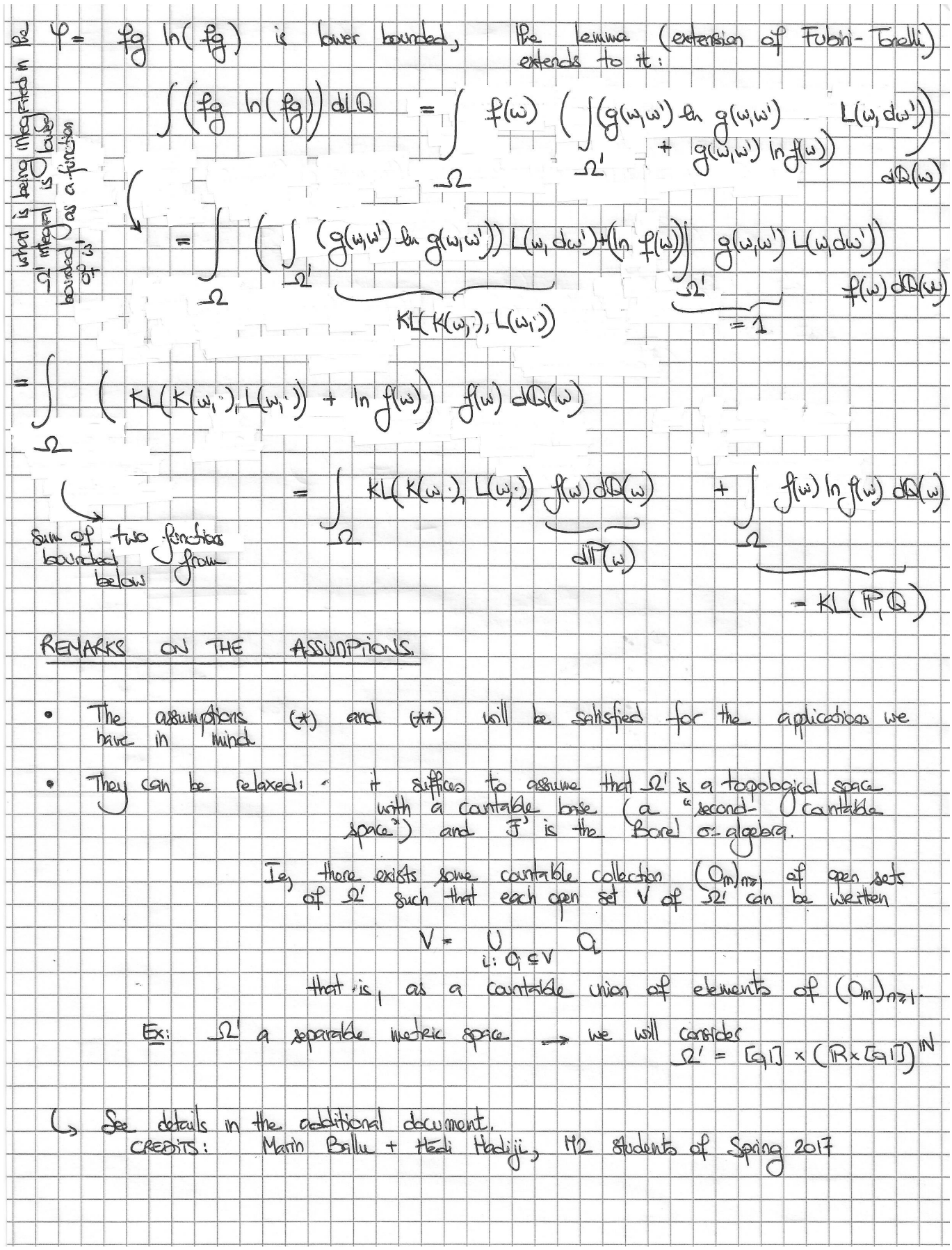
since $f = d\mathbb{P}/d\mathbb{Q}$

By Radon-Nikodym's theorem: $\frac{d K\mathbb{P}}{d\mathbb{LQ}} = f\, g$ $\mathbb{LQ}$-a.s.

* It is easily seen that $K\mathbb{P} \ll \mathbb{LQ} \Rightarrow \mathbb{P} \ll \mathbb{Q}$ (in all cases, even without (*) and (**))

* Therefore, we have $K\mathbb{P} \ll \mathbb{LQ} \Leftrightarrow \mathbb{P} \ll \mathbb{Q}$ under (*) and (**), we thus assume with no loss of generality that $K\mathbb{P} \ll \mathbb{LQ}$ and $\mathbb{P} \ll \mathbb{Q}$ (otherwise, both $= +\infty$ and the putative equality is $+\infty = +\infty$).

Then, $KL(K\mathbb{P}, \mathbb{LQ}) = \int_{\Omega\times\Omega'} \left(f(\omega)\, g(\omega,\omega') \ln f(\omega)\, g(\omega,\omega') \right) d\,\mathbb{LQ}(\omega,\omega')$



REMARKS ON THE ASSUMPTIONS

- The assumptions (*) and (**) will be satisfied for the applications we have in mind.
- They can be relaxed: it suffices to assume that Ω' is a topological space with a countable base (a "second-countable space") and $\mathcal{F}'$ is the Borel σ-algebra.

  I.e., there exists some countable collection $(O_m)_{m\geq 1}$ of open sets of Ω' such that each open set V of Ω' can be written

  $$V = \bigcup_{i : O_i \subseteq V} O_i$$

  that is, as a countable union of elements of $(O_m)_{m\geq 1}$.

  Ex: Ω' a separable metric space $\rightarrow$ we will consider $\Omega' = [0,1] \times (\mathbb{R} \times [0,1])^{\mathbb{N}}$

↳ See details in the additional document.
CREDITS: Matin Ballu + Hedi Hadiji, M2 students of Spring 2017

Part 2: Regret lower bound for stochastic bandits

Lower bounds on the regret for stochastic bandits.

Here is first a summary of the setting and context of stochastic bandits:

- K arms each indexed by $a = 1, 2, \dots K$
- With each arm is associated a probability distribution $\nu_a \in \mathcal{D}$
- $\mathcal{D}$ is the bandit model; a subset of $\mathcal{M}_1(\mathbb{R})$, the set of probability distributions over $\mathbb{R}$ with an expectation
- A bandit problem is denoted by $\nu = (\nu_a)_{a \in \{1,\dots K\}}$
- Important quantities and notation:

  $\mu_a = \mathbb{E}(\nu_a)$ is the expectation of ν_a

  $\mu^* = \max_{a=1,\dots K} \mu_a$ is the largest expectation within ν

  $\Delta_a = \mu^* - \mu_a$ is the gap for arm a

  Arm a is suboptimal if $\Delta_a > 0$

- Protocol: at each round $t = 1, 2, \dots$

  $U_0, U_1, U_2, \dots \overset{\text{iid}}{\sim} \mathcal{U}_{[0,1]}$

  1. The decision-maker picks $I_t \in \{1, \dots K\}$ possibly at random based on an auxiliary randomization U_{t-1}
  2. She gets a reward Y_t drawn at random according to ν_{I_t} (given I_t); this is the only piece of information she gets.

  (Note: the decision-maker knows the ν_a's but does not know the specific ν at hand)

- Aim / regret: maximize $\mathbb{E}\left[\sum_{t=1}^{T} Y_t\right]$

  which is equivalent to minimizing (controlling from above)

  $$R_T = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right]$$

- Rewriting by tower rule:

  $$R_T = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{I_t}\right] = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)]$$

  where $N_a(T) = \sum_{t=1}^{T} \mathbb{1}_{\{I_t = a\}}$ is the number of times arm a was pulled between 1 and T

⚠ It is thus necessary and sufficient to control $\mathbb{E}[N_a(T)]$ for suboptimal arms a

- What is a (randomized) strategy?

A sequence of measurable functions $(\psi_t)_{t\geq 0}$ with

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

H_t: history for the first t rounds; I_{t+1}: arm picked at round $t+1$

- Strategies that are consistent wrt a model $\mathcal{D}$:

If for all bandit problems $\nu \in \mathcal{D}^K$,

$$\forall \alpha \in (0,1], \quad \forall a \text{ s.t. } \Delta_a > 0, \quad \mathbb{E}[N_a(T)] = o(T^\alpha).$$

- Result: For "well behaved" models $\mathcal{D}$, there exist consistent strategies.

E.g.: at least $\mathcal{D} = \mathcal{M}_1([0,1])$, see the UCB strategy.

- Typical bounds for good strategies (stated in an asymptotic way, even though non-asymptotic bounds are available)

$$\forall \nu \in \mathcal{D}^K, \quad \forall a \text{ s.t. } \Delta_a > 0, \quad \limsup_{T\to+\infty} \frac{\mathbb{E}[N_a(T)]}{\ln T} \leq C_a(\nu)$$

where $C_a(\nu)$ is a problem-dependent constant.

- Optimal (in some sense) such constant: $C_a(\nu) = \dfrac{1}{\mathrm{Kinf}(\nu_a, \mu^*, \mathcal{D})} = \dfrac{1}{\mathrm{Kinf}(\nu_a, \mu^*)}$

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

with the convention: $\inf \emptyset = +\infty$.

We will only prove one part of this optimality: a lower bound on $C_a(\nu)$.

Theorem: (see Lai and Robbins, 1985; Burnetas and Katehakis, 1996)

For all bandit models $\mathcal{D} \subset \mathcal{M}_1(\mathbb{R})$,
For all strategies ψ consistent wrt $\mathcal{D}$ (possibly randomized),
For all bandit problems $\nu = (\nu_a)_{a\in\{1,\dots K\}} \in \mathcal{D}^K$,
For all suboptimal arms a (ie, such that $\Delta_a > 0$),

$$\liminf_{T\to+\infty} \frac{\mathbb{E}[N_a(T)]}{\ln T} \geq 1 / \mathrm{Kinf}(\nu_i, \mu^*, \mathcal{D})$$

Corollary: For all bandit models $\mathcal{D} \subseteq \mathcal{M}_1(\mathbb{R})$,
For all (possibly randomized) strategies ψ consistent wrt $\mathcal{D}$,
For all bandit problems $\nu = (\nu_a)_{a \in \{1..K\}} \in \mathcal{D}^K$,

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

To prove this theorem (and to prove other lower bounds), we will need the following fundamental inequality. In its statement, $\mathbb{P}_\nu$ and $\mathbb{E}_\nu$ refer to the probability distribution and the expectation induced by the bandit problem $\nu \in \mathcal{D}^K$.

Example: $\mathbb{P}_\nu^{H_T}$ is the law of $H_T = (U_0, Y_1, U_1 \dots Y_T, U_T)$ when the bandit problem is ν. Actually, $\mathbb{P}_\nu^{H_T}$ strongly depends on the strategy ψ used but we omit this dependency in the notation.

Lemma (Fundamental inequality for stochastic bandits):
For all bandit problems $\nu = (\nu_a)_{a \in \{1..K\}}$ and $\nu' = (\nu'_a)_{a \in \{1..K\}}$ in $\mathcal{D}^K$ with $\nu_a \ll \nu'_a$ for all a,
For all strategies
For all random variables Z taking values in $[0,1]$ and that are $\sigma(H_T)$-measurable,

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

The dependence on the strategy is hidden in the $\mathbb{E}_\nu[N_a(T)]$, $\mathbb{E}_\nu[Z]$ and $\mathbb{E}_{\nu'}[Z]$

Note: This lemma is our key to perform an IMPLICIT change of measures in the proof of the theorem.

Proof of the theorem (based on the lemma)

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

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

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

This is why we will

- Fix $\mathfrak{D}$, Ψ, $\vec{\nu}$ and a s.t. $\Delta_a > 0$
- Fix an alternative model $\vec{\nu}'$ of the form

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

That is, $\vec{\nu}$ and $\vec{\nu}'$ only differ at a; a is the unique optimal arm in $\vec{\nu}'$

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

Our fundamental inequality (the lemma) yields, since $\vec{\nu}$ and $\vec{\nu}'$ only differ at a:

$$E_{\vec{\nu}}[N_a(T)]\, KL(\vec{\nu}_a, \vec{\nu}'_a) \geqslant KL\left(\mathrm{Ber}\left(E_{\vec{\nu}}[N_a(T)/T]\right), \mathrm{Ber}\left(E_{\vec{\nu}'}\left[\frac{N_a(T)}{T}\right]\right)\right)$$

$$\geqslant -\ln 2 + \left(1 - E_{\vec{\nu}}[N_a(T)/T]\right) \ln \frac{1}{1 - E_{\vec{\nu}'}[N_a(T)/T]}$$

indeed: $KL(\mathrm{Ber}(p), \mathrm{Ber}(q))$

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

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

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

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 $\vec{\nu}$, a is suboptimal: $E_{\vec{\nu}}[N_a(T)/T] \longrightarrow 0$

- in the problem ν', all arms $k \neq a$ are suboptimal:

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

  ↳ in particular, for T large enough,

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

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

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

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

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

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

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

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 $1/\mathrm{KL}(\nu_a, \mu^*)$ lower bound.

Proof of the lemma:

- The inequality $\geq$ is a direct application of the data-processing inequality with expectations

- For the equality:

(1) We show by induction that $P_{\nu,\pi}^{H_T} = K_T\left(K_{T-1}\left(\cdots\left(K_1 m\right)\cdots\right)\right)$ (and similarly $P_{\nu,\pi}^{H_T} = K'_T\left(\cdots\left(K'_1 m\right)\cdots\right)$)

where K_t is the transition kernel:

$$h \in [0,1] \times (\mathbb{R}\times[0,1])^{t-1} \mapsto K_t(h,\cdot) = \nu_{\psi_{t-1}(h)} \otimes m$$

(we check below that π is regular)

Indeed: $T=0$: $H_0 = U_0 \sim \mathcal{U}_{[0,1]}$: $P_{\nu,\pi}^{U_0} = m$

$t \to t+1$: $\forall A \in \mathcal{B}([0,1]\times(\mathbb{R}\times[0,1])^t)$, $\forall B' \in \mathcal{B}(\mathbb{R})$, $\forall B \in \mathcal{B}([0,1])$,

$$P_{\nu,\pi}^{H_{t+1}}(A\times B'\times B) = P_{\nu,\pi}(H_t \in A \text{ and } Y_{t+1}\in B' \text{ and } U_{t+1}\in B)$$

$$= E_{\nu,\pi}\left[\mathbb{1}_A(H_t)\, P_{\nu,\pi}(Y_{t+1}\in B' \text{ and } U_{t+1}\in B \mid H_t)\right] \quad \text{(tower rule)}$$

$$= E_{\nu,\pi}\left[\mathbb{1}_A(H_t)\, \nu_{\psi_t(H_t)}(B')\, m(B)\right] \quad \text{(definition of the bandit model and of the strategy)}$$

$$= E_{\nu,\pi}\left[\mathbb{1}_A(H_t)\, K_{t+1}(H_t, B'\times B)\right] \quad \text{(definition of } K_{t+1}\text{)}$$

$$= \int \mathbb{1}_A\, K_{t+1}(h, B'\times B)\, dP_{\nu,\pi}^{H_t}(h) \quad \text{(more rewriting)}$$

$$= K_{t+1} P_{\nu,\pi}^{H_t}(A\times B'\times B) \quad \text{(definition of } K_{t+1} P_{\nu,\pi}^{H_t}\text{)}$$

(2) We first check that the assumptions of the chain rule are satisfied:

* The K_t are regular transition kernels: $\forall E \in \mathcal{B}(\mathbb{R})\otimes\mathcal{B}([0,1])$,

$$h \mapsto K_t(h,E) = \sum_{a=1}^{K} \mathbb{1}_{\{\psi_t(h)=a\}}\, \nu_a \otimes m(E)$$

is measurable as ψ_t is measurable

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

* Assumption (**): $(h, (y,u)) \longmapsto \frac{dK_t(h,\cdot)}{dK'_t(h,\cdot)}(y,u)$

is indeed bi-measurable
(product of measurable functions)

$$= \sum_{a=1}^{K} \mathbb{1}_{\{\psi_t(h)=a\}} \frac{d\nu_a}{d\nu'_a}(y)$$

(3) We then may apply the chain rule and show by induction the desired result based on:

- $KL(\mathbb{P}_{\nu}^{H_0}, \mathbb{P}_{\nu'}^{H_0}) = KL(\eta, \eta) = 0$
- For $t \geq 0$,

$$KL(\mathbb{P}_{\nu}^{H_{t+1}}, \mathbb{P}_{\nu'}^{H_{t+1}}) = KL(K_{t+1}\mathbb{P}_{\nu}^{H_t}, K'_{t+1}\mathbb{P}_{\nu'}^{H_t})$$

$$= KL(\mathbb{P}_{\nu}^{H_t}, \mathbb{P}_{\nu'}^{H_t}) + \int KL(K_{t+1}(h,\cdot), K'_t(h,\cdot))\, d\mathbb{P}_{\nu}^{H_t}(h)$$

$$= KL(\mathbb{P}_{\nu}^{H_t}, \mathbb{P}_{\nu'}^{H_t}) + \int KL(\nu_{\psi_t(h)} \otimes m, \nu'_{\psi_t(h)} \otimes m)\, d\mathbb{P}_{\nu}^{H_t}(h)$$

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

$$\mathbb{E}[\mathbb{1}_{\{\psi_t(H_t)=a\}}]$$

$$= \mathbb{E}[\mathbb{1}_{\{I_{t+1}=a\}}]$$

Exercise: $1/\mathrm{Kinf}(\vec{\nu}_a, \mu^*, \mathcal{D})$ vs. $8/\Delta_a^2$ for UCB

Recall that in the model $\mathcal{D} = \mathcal{P}([0,1])$, the UCB algorithm enjoys the following performance bound:

$$\forall \vec{\nu} \in \mathcal{P}([0,1])^K, \quad \forall a \text{ s.t. } \Delta_a > 0,$$
$$\mathbb{E}_{\vec{\nu}}[N_a(T)] \leq \frac{8}{\Delta_a^2} \ln T + 2.$$

Actually, there are refinements of UCB that get the distribution-dependent constant $8/\Delta_a^2$ arbitrarily close to $2/\Delta_a^2$.

But how do these $8/\Delta_a^2$ and $2/\Delta_a^2$ constants compare to $1/\mathrm{Kinf}(\vec{\nu}_a, \mu^*, \mathcal{P}([0,1]))$?

(1) For $p, q \in [0,1]$, we denote

$$\mathrm{kl}(p,q) = \mathrm{KL}(\mathrm{Ber}(p), \mathrm{Ber}(q))$$

Show that $\forall (p,q) \in [0,1]^2, \quad \mathrm{kl}(p,q) \geq 2(p-q)^2$.

(2) Show Pinsker's inequality: let $(\Omega, \mathcal{F})$ be a measurable space, let $\mathbb{P}, \mathbb{Q}$ be two distributions over $(\Omega, \mathcal{F})$, then:

$$\|\mathbb{P} - \mathbb{Q}\|_{TV} = \sup_{A \in \mathcal{F}} |\mathbb{P}(A) - \mathbb{Q}(A)| \leq \sqrt{\frac{1}{2}\mathrm{KL}(\mathbb{P}, \mathbb{Q})}$$

↑ the total variation distance between $\mathbb{P}$ and $\mathbb{Q}$

Even better, show the stronger form: $\sup_{Z \mathcal{F}\text{-measurable taking values in } [0,1]} |\mathbb{E}_{\mathbb{P}}[Z] - \mathbb{E}_{\mathbb{Q}}[Z]| \leq \sqrt{\frac{1}{2}\mathrm{KL}(\mathbb{P}, \mathbb{Q})}$

(3) Exhibit a lower bound on $\mathrm{Kinf}(\vec{\nu}_a, \mu^*, \mathcal{P}([0,1]))$ and conclude that some work is needed to get an upper bound matching our lower bound!

Exercice : Finite-time lower bound for small values of T

"All algorithms explore much!" ↪ We want to model that all algorithms must first explore uniformly all arms ($\leftrightarrow$ exploration) at least half of the time, before being able to perform exploitation more often.

(1) Establish the following local version of Pinsker's inequality:

- $\forall\, 0 \leq p < q \leq 1$, $\quad kl(p,q) \geqslant \dfrac{1}{2 \max\limits_{x \in [p,q]} x(1-x)} (p-q)^2$

$$\geqslant \frac{1}{2q}(p-q)^2$$

- Why is it stronger than the global version of Pinsker's inequality?

(2) Show that all strategies smaller than the uniform strategy [i.e. such that for all bandit problems, $\forall a$ s.t. $\mu_a = \mu^*$, $\mathbb{E}[N_a(T)] \geqslant T/K$], we have:

$$\forall\, T \leqslant \frac{1}{8\, KL^*},$$

where $KL^* = \max\limits_{j:\, \Delta_j > 0} Kinf(\nu_j, \mu_j^*)$

$\forall j$ s.t. $\Delta_j > 0$,

$$\mathbb{E}[N_j(T)] \geqslant \frac{1}{2}\,\frac{T}{K}$$

$\frac{1}{2}$: at least half of the time; $\frac{T}{K}$: uniform exploration

Hint: Consider the same alternative bandit problems ν' as in the theorem giving the asymptotic lower bound.