Stochastic bandits: setting and UCB strategy

Stochastic bandits. Finitely many arms.

Setting:

K arms indexed by $1, 2, \dots K$

With each arm j is associated a probability distribution ν_j (over $\mathbb{R}$) with an expectation

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

- The decision-maker picks $I_t \in \{1, \dots K\}$, possibly at random
- She gets a reward Y_t drawn at random according to ν_{I_t} (given I_t)
- This is the only feedback she gets / the only observation she has access to.

Aim:

We denote by $\mu_i = E(\nu_i)$ the expectation of ν_i

(note: operator E vs. expectation $\mathbb{E}$ of an expression involving random variables.)

Pseudo-regret $\bar{R}_T = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right]$ to be controlled

where $\mu^* = \max_{j \le K} \mu_j$

Useful notation:

$\Delta_a = \mu^* - \mu_a$ gap of arm a

$\Delta_a = 0$: a is an optimal arm (there can be several of them)

$\Delta_a > 0$: a is a suboptimal arm

$N_a(T) = \sum_{t=1}^{T} \mathbb{1}_{\{I_t = a\}}$ total number of times that a is pulled.

Note:

* Pseudo regret $\bar{R}_T$ is a very "expected" notion of regret

$\bar{R}_T \le$ probably $< \mathbb{E}\left[\max_{a=1,\dots K} \sum_{t}^{T} \dots - \sum_{t=1}^{T} Y_t\right]$

* Can be rewritten (see later) as $\bar{R}_T = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)]$

Upper confidence bound [UCB] algorithm: very popular!

For $t = 1, 2, \dots K$

- Pull arm $I_t = t$, get a reward Y_t

For $t = K+1, K+2, \dots$

- Pull an arm $I_t \in \operatorname{argmax}_{j \in \{1, \dots K\}} \left\{ \hat{\mu}_{j,t-1} + \sqrt{\frac{2 \ln t}{N_j(t-1)}} \right\}$

(the breaking rule: pick the element with smallest index)

where $N_j(t-1) = \sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = j\}}$

and where $\hat{\mu}_{j,t-1} = \frac{1}{N_j(t-1)} \sum_{s=1}^{t-1} Y_s \mathbb{1}_{\{I_s = j\}}$

always ≥ 1 since each arm was tried sequentially during rounds $1, 2, \dots K$

- Get a reward Y_t

Theorem: If the distributions ν_j have supports all included in $[0,1]$, then the pseudo-regret of UCB is smaller than

$$\overline{R}_T \leq \sum_{i : \Delta_i > 0} \left(\frac{8 \ln T}{\Delta_i} + 2 \right)$$

This regret bound is obtained via the following proposition:

Proposition: If the distributions ν_j have supports all included in $[0,1]$, then

$$\forall i \text{ s.t. } \Delta_i > 0, \quad \mathbb{E}[N_i(T)] \leq \frac{8 \ln T}{\Delta_i^2} + 2.$$

Exercise The bounds above are called distribution-dependent because they depend heavily on the distributions ν_i at hand (via the gaps $\Delta_i = \mu^* - \mu_i$).
Show the following distribution-free bound (that only

depends on the support $[a,1]$, not on the specific distributions ν_i at hand) : for the UCB algorithm,

$$\sup_{\nu_1,\dots\nu_K \text{ with supports in } [a,1]} \bar{R}_T \leq O\left(\sqrt{TK\ln T}\right).$$

Hint: For small values of Δ_i, the bound of the Proposition can be worse than the trivial T bound...

Proof [of the theorem based on the Proposition] :

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

where by definition of the bandit model, $\mathbb{E}[Y_t \mid I_t] = \mu_{I_t}$ ← Given I_t, Y_t is drawn at random according to ν_{I_t}

thus (by the tower rule) $\mathbb{E}[Y_t] = \mathbb{E}[\mu_{I_t}]$

$$= \sum_j \mu_j \, \mathbb{E}\left[\mathbb{1}_{\{I_t=j\}}\right]$$

Summing over t:

$$\mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] = \sum_{j=1}^{K} \mu_j \, \mathbb{E}[N_j(T)]$$

and (in view of $T = \sum_j \mathbb{E}[N_j(T)]$)

$$\bar{R}_T = \sum_j (\mu^* - \mu_j)\,\mathbb{E}[N_j(T)] = \sum_{j=1}^{K} \Delta_j \, \mathbb{E}[N_j(T)]$$

$$= \sum_{j:\Delta_j>0} \Delta_j \, \mathbb{E}[N_j(T)]$$

↓ it suffices to consider the suboptimal arms...

We conclude by substituting $\mathbb{E}[N_j(T)] \leq \frac{8\ln T}{\Delta_j^2} + 2$ and by bounding $2\Delta_j \leq 2$.

NOTE: Keep in mind the rewriting as we will often use it!

$$\bar{R}_T = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] = \sum_{a=1}^{K} \Delta_a \, \mathbb{E}[N_a(T)]$$

Proof [of the Proposition]: We fix an optimal arm $a^* \in \{1, \dots K\}$, i.e. s.t. $\mu_{a^*} = \mu^*$.

→ It will show why this algorithm is called UCB:

Because $\hat{\mu}_{j,t-1} + \sqrt{\frac{2 \ln t}{N_j(t-1)}}$ will indeed appear as an upper confidence bound on μ_j

- $\hat{\mu}_{j,t-1}$: estimate based on the raw performance ↔ exploitation of the results
- $\sqrt{\frac{2 \ln t}{N_j(t-1)}}$: larger for arms not much sampled so far ↔ forces some exploration

The UCB algorithm realizes some compromise / trade-off between exploitation & exploration.

Later on we compare these statements to the Hoeffding-Azuma inequality [cf. Lecture 3]

→ **LEMMA:** $\forall j, \forall t \geq j$ (so that $N_j(t) \geq 1$)

$$\forall \delta \in (0,1), \quad \mathbb{P}\left\{ \mu_j > \hat{\mu}_{j,t} - \sqrt{\frac{\ln(1/\delta)}{2 N_j(t)}} \right\} \geq 1 - t\delta.$$

replace X_i by $1 - X_i$ then μ_a becomes $1 - \mu_a$ and $\hat{\mu}_{a,t}$ becomes $1 - \hat{\mu}_{a,t}$, but will be supported in $[0,1]$ as well

→ By symmetry or:

$$\forall \delta \in (0,1), \quad \mathbb{P}\left\{ \mu_j < \hat{\mu}_{j,t} + \sqrt{\frac{\ln(1/\delta)}{2 N_j(t)}} \right\} \geq 1 - t\delta$$

→ Application:

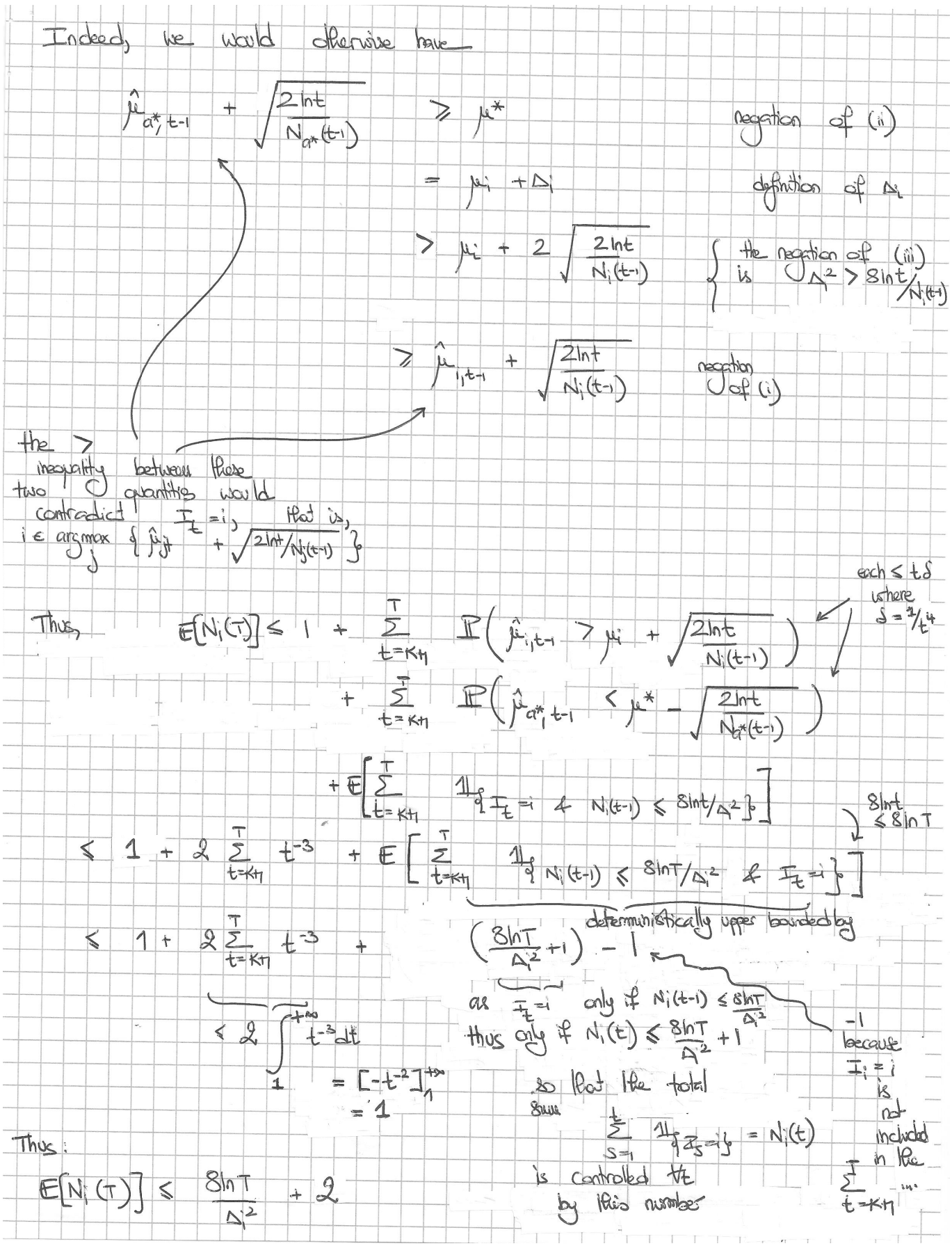
$$N_i(T) = 1 + \sum_{t=K+1}^{T} \mathbb{1}_{\{I_t = i\}}$$

We show below that $t \geq K+1$ and $I_t = i$ entails one of the following:

(i) $\hat{\mu}_{i,t-1} > \mu_i + \sqrt{\frac{2 \ln t}{N_i(t-1)}}$ [$\mu_i <$ lower confidence bound]

(ii) $\hat{\mu}_{a^*,t-1} < \mu^* - \sqrt{\frac{2 \ln t}{N_{a^*}(t-1)}}$ [$\mu^* >$ upper confidence bound]

(iii) $N_i(t-1) \leq \frac{8 \ln t}{\Delta^2}$ [i not played often enough yet]


Indeed, we would otherwise have
negation of (ii)
definition of Δ_i
the negation of (iii) is $\Delta_i^2 > 8\ln t / N_i(t-1)$
negation of (i)
the > inequality between these two quantities would contradict $I_t = i$, that is, $i \in \arg\max_j \{\hat{\mu}_{j,t} + \sqrt{2\ln t / N_j(t-1)}\}$
Thus, $\mathbb{E}[N_i(T)] \le 1 + \sum_{t=K+1}^{T} \mathbb{P}\left(\hat{\mu}_{i,t-1} > \mu_i + \sqrt{\frac{2\ln t}{N_i(t-1)}}\right)$
each $\le t\delta$ where $\delta = 1/t^4$
$+ \sum_{t=K+1}^{T} \mathbb{P}\left(\hat{\mu}_{a_i^*,t-1} < \mu^* - \sqrt{\frac{2\ln t}{N_{a_i^*}(t-1)}}\right)$
$+ \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}\{I_t = i \text{ et } N_i(t-1) \le 8\ln t/\Delta_i^2\}\right]$
$8\ln t \le 8\ln T$
$\le 1 + 2\sum_{t=K+1}^{T} t^{-3} + \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}\{N_i(t-1) \le 8\ln T/\Delta_i^2 \text{ et } I_t = i\}\right]$
deterministically upper bounded by
$\le 1 + 2\sum_{t=K+1}^{T} t^{-3} + \left(\frac{8\ln T}{\Delta_i^2} + 1\right) - 1$
as $I_t = i$ only if $N_i(t-1) \le \frac{8\ln T}{\Delta_i^2}$
thus only if $N_i(t) \le \frac{8\ln T}{\Delta_i^2} + 1$
so that the total $\sum_{s=1}^{t} \mathbb{1}\{I_s = i\} = N_i(t)$ is controlled the by this number
-1 because $I_t = i$ is not included in the $\sum_{t=K+1}^{T}$
$\le 2\int_1^{+\infty} t^{-3}\,dt = [-t^{-2}]_1^{+\infty} = 1$
Thus:
$\mathbb{E}[N_i(T)] \le \frac{8\ln T}{\Delta_i^2} + 2$

Let's complete the proof of the Lemma: ["Hoeffding-Azuma inequality with a random number of summands"]

Setting: Probability distributions $\nu_1, \dots \nu_K$ over $[0,1]$
with respective expectations $\mu_1, \dots \mu_K$

At each round, $I_t \in \{1, \dots K\}$ is picked in a $\sigma(Y_1, \dots Y_{t-1})$-measurable way

then Y_t is drawn independently at random according to ν_{I_t}, given I_t

ie: $Y_t \mid I_t \sim \nu_{I_t}$

We denote $N_a(t) = \sum_{s=1}^{t} \mathbb{1}_{\{I_s = a\}}$ and assume that each arm a was pulled once in the first K rounds,

so that:

$$N_a(t) \geq 1 \quad \forall t \geq K$$

Then, for $t \geq K$:

$$\hat{\mu}_{a,t} = \frac{1}{N_a(t)} \sum_{s=1}^{t} Y_s \mathbb{1}_{\{I_s = a\}}$$

Lemma: $\forall \delta \in (0,1)$,

$$\mathbb{P}\left\{ \mu_a > \hat{\mu}_{a,t} - \sqrt{\frac{\ln(1/\delta)}{2 N_a(t)}} \right\} \geq 1 - t\delta$$

(and a similar symmetric statement with $\mu_a < \hat{\mu}_{a,t} + \sqrt{\dots}$)

The proof will be based on the fact that $(Z_t)_{t \geq 0}$, where

$$Z_t = \sum_{s=1}^{t} (Y_s - \mu_a) \mathbb{1}_{\{I_s = a\}}$$

is a martingale wrt. $(\mathcal{F}_t) = (\sigma(Y_1, \dots Y_t))_{t \geq 0}$

$$\mathbb{E}\left[(Y_t - \mu_a) \mathbb{1}_{\{I_t = a\}} \mid Y_1 \dots Y_{t-1} \right] = \mathbb{E}\left[(Y_t - \mu_a) \mathbb{1}_{\{I_t = a\}} \mid I_t, Y_1 \dots Y_{t-1} \right]$$

$$= \left(\mathbb{E}[Y_t \mid I_t, Y_1 \dots Y_{t-1}] - \mu_a \right) \mathbb{1}_{\{I_t = a\}}$$

(where we used the bandit model)

$$= (\mu_{I_t} - \mu_a) \mathbb{1}_{\{I_t = a\}} = 0 \quad \text{a.s.}$$

Remark: How does this bound compare to what the classical version of the Hoeffding-Azuma says?

Martingale increment $(Y_s - \mu_a) \mathbb{1}_{\{I_s = a\}}$ bounded between

$$a_t = -\mu_a \text{ and } b_t = 1 - \mu_a$$

so that $(b_t - a_t)^2 = 1$

(actually in the version I stated, I can take $\leq$ or $<$)

$$1-t\delta \leq \mathbb{P}\left\{ Z_t < \sqrt{\frac{t}{2}\ln\frac{1}{t\delta}} \right\} = \mathbb{P}\left\{ N_a(t)\left(\hat{\mu}_{a,t} - \mu_a\right) < \sqrt{\frac{t}{2}\ln\frac{1}{t\delta}} \right\}$$

$$= \mathbb{P}\left\{ \hat{\mu}_{a,t} - \sqrt{\frac{t}{N_a(t)}}\sqrt{\frac{\ln(1/t\delta)}{2\,N_a(t)}} < \mu_a \right\}$$

versus the bound of our lemma: $1-t\delta \leq \mathbb{P}\left\{ \hat{\mu}_{a,t} - \sqrt{\frac{\ln(1/\delta)}{2N_a(t)}} < \mu_a \right\}$

The proposed deviations essentially differ from a $\sqrt{t/N_a(t)}$ factor, and it is so nice to get rid of it!

Proof: (1) We prove that $\forall x \in \mathbb{R}$, $\mathbb{E}\left[e^{xZ_t - x^2/8\, N_a(t)} \right] \leq 1$

We do so by showing that $M_t = \exp\left(xZ_t - \frac{x^2}{8} N_a(t) \right)$ is a supermartingale, so that $\mathbb{E}[M_t] \leq \mathbb{E}[M_0] = 1$.

Indeed, by the conditional version of Hoeffding's lemma,

$$\mathbb{E}\left[e^{x(Y_t - \mu_a)\mathbb{1}_{\{I_t = a\}}} \,\middle|\, \mathcal{F}_{t-1} \right] \leq e^{x^2/8} \quad \text{a.s.}$$

but we can do better!

Since I_t and thus also $\mathbb{1}_{\{I_t=a\}}$ are $\mathcal{F}_{t-1}$ - measurable, we get:

$$\mathbb{E}\left[e^{x(Y_t-\mu_a)\mathbb{1}_{\{I_t=a\}}} \,\middle|\, \mathcal{F}_{t-1} \right] = \mathbb{E}\left[e^{x(Y_t-\mu_a)\mathbb{1}_{\{I_t=a\}}} \left(\mathbb{1}_{\{I_t=a\}} + \mathbb{1}_{\{I_t\neq a\}} \right) \,\middle|\, \mathcal{F}_{t-1} \right]$$

$$= \mathbb{E}\left[e^{x(Y_t-\mu_a)\mathbb{1}_{\{I_t=a\}}} \,\middle|\, \mathcal{F}_{t-1} \right] \mathbb{1}_{\{I_t=a\}} + e^0\, \mathbb{1}_{\{I_t\neq a\}}$$

(given what we had before)

$$\leq e^{x^2/8}\, \mathbb{1}_{\{I_t=a\}} + \mathbb{1}_{\{I_t\neq a\}} = \exp\left(\frac{x^2}{8}\, \mathbb{1}_{\{I_t=a\}} \right)$$

Put differently, $\mathbb{E}\left[e^{x(Y_t-\mu_a)\mathbb{1}_{\{I_t=a\}} - \frac{x^2}{8}\mathbb{1}_{\{I_t=a\}}} \,\middle|\, \mathcal{F}_{t-1} \right] \leq 1$

which entails that $\exp\left(x \sum_{s=1}^{t} (Y_s - \mu_a)\mathbb{1}_{\{I_s=a\}} - \frac{x^2}{8} \sum_{s=1}^{t} \mathbb{1}_{\{I_s=a\}} \right)$

$$= \exp\left(xZ_t - \frac{x^2}{8} N_a(t) \right) = M_t$$

is a supermartingale wrt $\mathcal{F}_t = \sigma(Y_1 \dots Y_t)$.

(2) We prove that $\forall \varepsilon > 0$, $\forall \ell \geq 1$, $\mathbb{P}\{ Z_t \geq \varepsilon \text{ and } N_a(t) = \ell \} \leq \exp(-2\varepsilon^2/\ell)$

Indeed, by a Markov-Chernoff bounding, $\forall x > 0$,

$$\mathbb{P}\{ Z_t \geq \varepsilon \text{ and } N_a(t) = \ell \} \leq e^{-x\varepsilon} \mathbb{E}\left[e^{x Z_t} \mathbb{1}_{\{N_a(t)=\ell\}} \right]$$

$$= e^{-x\varepsilon + \frac{x^2 \ell}{8}} \mathbb{E}\left[e^{x Z_t - \frac{x^2}{8} N_a(t)} \mathbb{1}_{\{N_a(t)=\ell\}} \right]$$

$$\leq e^{-x\varepsilon + x^2\ell/8} \underbrace{\mathbb{E}\left[e^{x Z_t - \frac{x^2}{8} N_a(t)} \right]}_{\leq 1 \text{ by (1)}}$$

Optimizing over $x > 0$ (take $x = 4\varepsilon/\ell$) yields the claimed bound.

(3) Conclusion: we prove that $\mathbb{P}\left\{ \mu_a \leq \hat{\mu}_{at} - \sqrt{\frac{\ln(1/\delta)}{2N_a(t)}} \right\} \leq t\delta$

Indeed, by distinguishing according to the values taken by $N_a(t)$:

$$\mathbb{P}\left\{ \mu_a \leq \hat{\mu}_{at} - \sqrt{\frac{\ln(1/\delta)}{2N_a(t)}} \right\}$$

$$\leq \sum_{\ell=1}^{t} \mathbb{P}\left\{ N_a(t) = \ell \text{ and } \mu_a \leq \hat{\mu}_{at} - \sqrt{\ln(1/\delta)/2\ell} \right\}$$

$$= \sum_{\ell=1}^{t} \mathbb{P}\left\{ N_a(t) = \ell \text{ and } \frac{Z_t}{N_a(t)} \geq \sqrt{\ln(1/\delta)/2\ell} \right\}$$

$$= \sum_{\ell=1}^{t} \mathbb{P}\left\{ N_a(t) = \ell \text{ and } Z_t \geq \sqrt{\ell \ln(1/\delta)/2} \right\}$$

by (2) $\leq \sum_{\ell=1}^{t} \exp\left(-2 \left(\ell \ln(1/\delta)/2 \right) / \ell \right) = t\delta.$

($\sum_{\ell=1}^{t-K+1}$ would be enough)

Two notes on this proof:

* We noted last week that the conditional version of Hoeffding's lemma (based on Proof #2) could be generalized into

X bounded random variable, U, V two $\mathcal{G}$-measurable random variables with $U \leq X \leq V$

then $\forall s \in \mathbb{R}$, $\ln \mathbb{E}[e^{sX} | \mathcal{G}] \leq s\,\mathbb{E}[X|\mathcal{G}] + \frac{s^2}{8}(V-U)^2$

(Because of the proof technique relying on weights with denominators $1/V-U$, it is safer to assume $V-U>0$ as, eg, by replacing U and V by $U-\varepsilon$ and $V+\varepsilon$ if needed and letting $\varepsilon \searrow 0$; so at the end of the day we may drop the $V-U>0$ as condition.)

* This extension may be applied to

$$X_t = (Y_t - \mu_a)\,\mathbb{1}_{\{I_t=a\}}$$
$$\mathcal{G} = \mathcal{F}_{t-1}$$
$$U_t = -\mu_a\,\mathbb{1}_{\{I_t=a\}}$$
$$V_t = (1-\mu_a)\,\mathbb{1}_{\{I_t=a\}}$$

and directly entails

$$\mathbb{E}\left[e^{s(Y_t-\mu_a)\mathbb{1}_{\{I_t=a\}}} \,\middle|\, \mathcal{F}_{t-1} \right] \leq \exp\left(\frac{s^2}{8}\,\mathbb{1}_{\{I_t=a\}} \right)$$

without the need for the

$$1 = \mathbb{1}_{\{I_t=a\}} + \mathbb{1}_{\{I_t\neq a\}}$$

trick used in Step (1).

* The question is: Don't we have a generalized version of the Hoeffding-Azuma inequality with such predictable ranges $V_t - U_t$?

Yes we do have something in terms of constant upper bounds $V_t - U_t \leq \Delta_t \in \mathbb{R}$ as

But $V_t - U_t = \mathbb{1}_{\{I_t=a\}}$ can only be bounded by $\Delta_t = 1$

So I think that Steps (2) and (3) are needed

Stochastic bandits with a continuum of arms

Stochastic bandits :

What about arms indexed by a continuum?

Setting 1 : Arms indexed by $x \in \mathcal{A}$, where $\mathcal{A}$ is some possibly large set

With each arm $x \in \mathcal{A}$ is associated a probability distribution ν_x over $\mathbb{R}$ s.t. $\mathbb{E}(\nu_x)$ exists

At each round, the decision-maker picks $I_t \in \mathcal{A}$, gets a reward Y_t drawn at random according to ν_{I_t} (given I_t); and this is the only feedback she gets.

Definition : $f : x \in \mathcal{A} \longmapsto \mathbb{E}(\nu_x)$ is the mean-payoff function

Regret : $\overline{R}_T = T \sup_{x \in \mathcal{A}} f(x) - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right]$

Setting 2 : [special case] $\longrightarrow$ Noisy optimization of a function.

We fix $f : \mathcal{A} \to \mathbb{R}$

The noise is given by a sequence of iid random variables $\varepsilon_1, \varepsilon_2, \dots$

When $I_t \in \mathcal{A}$ is picked, $Y_t = f(I_t) + \varepsilon_t$

↳ Special case of setting #1 where ν_x is the distribution of $f(x) + \varepsilon_1$ (all these distributions have the same shape given by the common distribution of the ε_j)

We of course need conditions for the regret to be minimized.

Definition : Let $\mathcal{F}$ be a set of possible bandit problems $\nu = (\nu_x)_{x \in \mathcal{A}}$

→ The regret can be controlled (in a non-uniform way) against $\mathcal{F}$ if:

there exists a strategy s.t. $\forall \nu \in \mathcal{F}$, $\overline{R}_T = o(T)$.

we also say that $(\mathcal{A}, \mathcal{F})$ is tractable

Ex: $\mathcal{A} = \{1, \dots K\}$ and $\mathcal{F} = \big(\mathcal{P}([0,1])\big)^K$, the set of all K-tuples of probability distributions over $[0,1]$

the case of finitely many arms with bounded distributions

→ UCB does the job.

Counter-example: $\mathcal{A} = [0,1]$ and $\mathcal{F} = \big(\mathcal{P}([0,1])\big)^{[0,1]}$

↑ illustrating that continuity is a minimal requirement.

all bandit problems $(\nu_x)_{x\in[0,1]}$ with distributions ν_x having support $[0,1]$.

Indeed: Consider $(\delta_0)_{x\in[0,1]}$ the bandit problem in which each arm x is associated with the Dirac mass on 0

Fix any strategy: it gets $Y_t = 0$ $\forall t$ and uses a sequence of (possibly) random choices I_t, $t \geq 1$

Since probability distributions can only have at most countably many atoms, $\mathcal{Y} = \{x \in [0,1] : \exists t \mid \mathbb{P}[I_t = x] > 0 \text{ under } (\delta_0)_{x\in[0,1]}\}$ is countable. In particular, $[0,1] \setminus \mathcal{Y}$ is non empty.

But the strategy behaves the same under the problem $(\nu'_x)_{x\in[0,1]}$ in which $\begin{cases} \nu'_x = \delta_0 & \forall x \neq x_0 \\ \nu'_{x_0} = \delta_1 & \text{for one fixed } x_0 \in [0,1] \setminus \mathcal{Y} \end{cases}$

With probability 1, it thus never hits x_0.

Therefore, $Y_t = 0$ a.s. $\forall t$ and $\overline{R}_T = T - \mathbb{E}\big[\sum_{t=1}^{T} Y_t\big] = T.$

Actually, continuity is sufficient for the regret to be controlled, as long as $\mathcal{A}$ is not too large.

Theorem: Let $\mathcal{A}$ be a cont metric space and let $\mathcal{F}$ be the set of bandit problems $(\nu_x)_{x\in\mathcal{A}}$ with:
- → $\forall x$, ν_x is a distribution over $[0,1]$
- → a continuous mean-payoff function $f : x \mapsto \mathbb{E}(\nu_x)$

The regret can be controlled against $\mathcal{F}^{\text{fast}}$ if and only if $\mathcal{A}$ is separable.

Corollary. Let $\mathcal{F}^{\text{all}}$ be the family of all bandit models $(\nu_x)_{x\in\mathcal{A}}$ with distributions ν_x over $[0,1]$. Then the regret against $\mathcal{F}^{\text{all}}$ can be controlled if and only if $\mathcal{A}$ is at most countable.

(let $\mathcal{A}$ be any set.)

Before we prove these facts, consider the following more concrete example, in which, by strengthening the regularity requirement on the mean-payoff function, we can even get rates.

Exercise (Lipschitz bandits): Let $\mathcal{A}=[0,1]$ and let $\mathcal{F}^{\text{Lip}}$ be the family of bandit models $(\nu_x)_{x\in[0,1]}$ with distributions ν_x over $[0,1]$ and with mean-payoff functions that are Lipschitz.

Exhibit a strategy based on UCB + a sequence of discretizations of $[0,1]$ into K bins (to be refined over time) such that:

$$\forall \nu\in\mathcal{F}^{\text{Lip}},\quad \overline{R}_T \leq \left(3L + 6\sqrt{8\ln T} + 2\right)T^{2/3} + 2$$

where L is the Lipschitz constant of the mean-payoff function of ν

Hint: First, prove a performance bound by splitting $[0,1]$ into $\left[\frac{i-1}{K}, \frac{i}{K}\right]$ with $i=1,\dots,K$ for a fixed K, where each bin $\left[\frac{i-1}{K}, \frac{i}{K}\right]$ plays the role of an i in a bandit problem with finitely many arms. Then discuss how to pick K over the time, as we do in the next proof.

Proof of the Corollary:

We endow $\mathcal{A}$ with the discrete topology, i.e., choose the distance $d(x,y) = \mathbb{1}\{x\neq y\}$. Then:

1. All applications $f:\mathcal{A}\to\mathbb{R}$ are continuous
2. $\mathcal{A}$ is separable if and only if $\mathcal{A}$ is at most countable.

Proof of the Theorem: It relies on the possibility or impossibility of uniform exploration of the arms

1) **If $\mathcal{A}$ is separable:** let $(x_n)_{n\in\mathbb{N}}$ be a collection of points in $\mathcal{A}$ that is dense

We pick actions in a triangular fashion:

Regime 1: UCB based on x_1 x_2 (fresh start) : $I_1^{(1)} \dots I_4^{(1)}$

⋮

Regime r: UCB based on $x_1 x_2 \dots x_r$ (fresh start) $I_1^{(r)} \dots I_{(r+1)^2}^{(r)}$

⋮

In Regime r: ↑ starts at time $S_r + 1 = 2^2 + \dots r^2 + 1$

$$(r+1)^2 \max_{s\le r} f(x_s) - \mathbb{E}\left[\sum_{t=S_r+1}^{S_r+(r+1)^2} Y_t\right] \qquad (*)$$

$$\le c\sqrt{r^3 \ln r}$$

↑ well-chosen numerical constant

↖ order of magnitude of the distribution-free regret bound for UCB on $(r+1)^2$ steps with r arms (we saw this bound as an exercise)

Now, let $\varepsilon > 0$ and let $\tilde{r}_\varepsilon \in \mathbb{N}^*$ s.t. $\mu_{x_{\tilde{r}_\varepsilon}} \ge \sup_{\mathcal{A}} f - \varepsilon$

($\tilde{r}_\varepsilon$ exists by separability of $\mathcal{A}$ and continuity of f)

In particular, $\max_{s\le \tilde{r}_\varepsilon} f(x_s) \ge \sup_{\mathcal{A}} f - \varepsilon \qquad (**)$

We denote by r_T the index of the regime where T lies:

we have that S_r is of the order of r^3

thus that r_T is of the order of $T^{1/3}$

in particular, $r_T = O(T^{1/3})$

The regret can be decomposed (for T large enough) as

$$\overline{R}_T = T \sup_{A} f - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] = \text{sum of the regrets of each regime}$$

$$= \underbrace{\sum_{r=1}^{\tilde{r}_\varepsilon - 1} (r+1)^2}_{\text{initial regimes, regret bounded by their lengths}} + \underbrace{\sum_{r=\tilde{r}_\varepsilon}^{r_T - 1} \left((r+1)^2 \varepsilon + c\sqrt{r^3 \ln r} \right)}_{\text{cf. bounds (*) and (**)}} + \underbrace{(r_T+1)^2}_{\text{regime } r_T \text{ may be incomplete}}$$

$$= O(1)$$

$$\leqslant T\varepsilon + c \sum_{r=\tilde{r}_\varepsilon}^{r_T - 1} r^{3/2}\sqrt{\ln r_T} = O(T^{2/3})$$

$$\leqslant T\varepsilon + O\left(r_T^{5/2}\sqrt{\ln r_T}\right)$$

$$= T\varepsilon + O\left(T^{5/6}\sqrt{\ln T}\right)$$

All in all, $\limsup \frac{\overline{R}_T}{T} \leqslant \varepsilon$, which is true $\forall \varepsilon > 0$,

that is, $\lim \frac{\overline{R}_T}{T} = 0$ as requested.

2) If $\mathcal{A}$ is not separable :

* We use the following characterization of separability (which relies on Zorn's lemma) :

> A metric space $\mathcal{X}$ is separable if and only if it contains no uncountable subset $\mathcal{D}$ s.t. $\rho = \inf\{d(x,y) : x,y \in \mathcal{D}\} > 0$.

In particular, if $\mathcal{A}$ is not separable, there exist an uncountable subset $\mathcal{D} \subset \mathcal{A}$ and $\rho > 0$ such that the balls $B(a, \rho/2)$, with $a \in \mathcal{D}$, are all disjoint.

↪ No probability distribution over $\mathcal{A}$ can give a positive mass to all these balls.

* We consider the bandit models $\nu^{(a)}$ inducing mean-payoff functions $f^{(a)} : x \in \mathcal{A} \longmapsto \left(1 - \frac{d(x,a)}{\rho/2}\right)^+$; in particular, $\nu_x^{(a)} = \delta_0$ for $x \notin B(a, \rho/2)$.

↑ $f^{(a)}$ is indeed continuous.

We proceed as in the example showing the necessity of continuity when $\mathcal{A} = (0,1]$ and consider the bandit model $(\delta_0)_{x \in \mathcal{A}}$, as well as any strategy and the laws induced by the I_t under this model : let ν_t be the law of I_t under $(\delta_0)_{x \in \mathcal{A}}$ and let $\nu = \sum_{t \geq 1} \frac{1}{2^t} \nu_t$.

[as only countably many balls can have a positive mass under ν]

↪ There exists $a \in \mathcal{A}$ s.t. $\nu(B(a, \rho/2)) = 0$, that is, s.t.,

$$\forall t \geq 1, \quad \mathbb{P}\left(I_t \in B(a, \rho/2) \text{ under } (\delta_0)_{x \in \mathcal{A}}\right) = 0.$$

The considered strategy is therefore such that the I_t have the same distribution under $(\delta_0)_{x \in \mathcal{A}}$ and $\nu^{(a)}$. In particular, $\mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] = 0$ in both cases, but in the latter case, $\sup f^{(a)} = 1$, so that $\overline{R}_T = T$ against $\nu^{(a)}$. The regret is not controlled against $\nu^{(a)} \in \mathcal{F}^{\text{cont}}$.

Tools for regret lower bounds for stochastic bandits

[Back to K-armed bandits]

Overview of the next topic: Fix a model $\mathcal{D}$, known to the decision-maker, ie, a collection of probability distributions over $\mathbb{R}$ with an expectation.

Assume that $\nu_1, \dots \nu_K$ are unknown but that the decision-maker knows $\nu_j \in \mathcal{D}$ $\forall j$.

What are the best bounds on $\overline{R}_T = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right]$?

One can show matching upper and lower bounds (with associated strategies):

$\overline{R}_T$ is at best of the order of $\left(\sum_{a:\Delta_a>0} \frac{\Delta_a}{\mathrm{Kinf}(\nu_a, \mu^*, \mathcal{D})}\right) \ln T$

where

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

Kullback-Leibler divergence

expectation of ν'_a

We will only prove the lower bound part but not discuss the strategy, called KL-UCB, to achieve the bound.

} it's just a matter of time, we would need about 3h to discuss this strategy

* But * before we do that, I guess that some reminder of basic and non-basic results about KL divergences would be needed!

The Kullback-Leibler divergence: definition and basic properties.

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

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

Basic facts:

- Existence of the defining integral when $\mathbb{P} \ll \mathbb{Q}$: because $\psi : x \mapsto x \ln x$ is bounded from below on $[0, +\infty)$
- $KL(\mathbb{P},\mathbb{Q}) \geq 0$ and $0 = KL(\mathbb{P},\mathbb{Q})$ if and only if $\mathbb{P} = \mathbb{Q}$:

It suffices to consider the case $\mathbb{P} \ll \mathbb{Q}$: because ψ is strictly convex, Jensen's inequality indicates that

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

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

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

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:

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

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

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

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

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

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

$$\overset{not.}{=} \int_\Omega \mathbb{1}_B(X)\, \gamma(x)\, d\mathbb{Q} = \int_{\Omega'} \mathbb{1}_B\, \gamma\, d\mathbb{Q}^X.$$

(by definition of $\mathbb{Q}^X$)

Therefore,

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

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

$$\leq \int_\Omega \mathbb{E}_{\mathbb{Q}}\left[\frac{d\mathbb{P}}{d\mathbb{Q}} \ln \frac{d\mathbb{P}}{d\mathbb{Q}} \,|\, X\right] d\mathbb{Q} \quad \text{(conditional version of Jensen's inequality)}$$

$$= \int_\Omega \left(\frac{d\mathbb{P}}{d\mathbb{Q}} \ln \frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q} = KL(\mathbb{P}, \mathbb{Q}) \quad \text{(tower rule)}$$

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 $(x,y) \in [0,+\infty)^2 \mapsto \left(\frac{x}{y} \ln \frac{x}{y}\right) y$

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 $\alpha \in (0,1)$,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

we just use that for $f\geq 0$ a constant, $\int f d\mu = \int f \mathbb{1}_A d\mu + \int f \mathbb{1}_{A^c} d\mu$ (whether f is integrable or not)

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

on $\Omega\times\{1\}$, $d\mathbb{Q}$ is $(1-\alpha)d\mathbb{Q}_1$

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

$$= \int_\Omega \left(\frac{d\mathbb{P}_1}{d\mathbb{Q}_1}(\omega)\ln\frac{d\mathbb{P}_1}{d\mathbb{Q}_1}(\omega)\right)(1-\alpha)\,d\mathbb{Q}_1(\omega) + \int \cdots$$

$= KL(\mathbb{P}_1,\mathbb{Q}_1)\times(1-\alpha)$ $\qquad$ $\alpha\,KL(\mathbb{P}_2,\mathbb{Q}_2)$

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:

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