We first complete the proof of the distribution-dependent lower bound.

Protocol: with randomized strategies $(\psi_t)_{t \geq 0}$, where

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

We have: $\begin{cases} Y_t \mid I_t \sim \nu_{I_t} & \forall t \\ U_0, U_1, \ldots \text{ iid} \sim \mathcal{U}_{[0,1]} \end{cases}$

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

by the tower rule

where $\mu_a = \mathbb{E}(\nu_a)$, $\mu^* = \max_{a=1\ldots K} \mu_a$,

$\Delta_a = \mu^* - \mu_a$ and $\mathbb{E}_\nu$ denotes the expectation when the underlying bandit problem is $\nu = (\nu_a)_{a \in \{1 \ldots K\}}$

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

Definitions:

- Bandit model $\mathcal{D}$: A subset of the set $\mathcal{M}_1(\mathbb{R})$ of all probability distributions over $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with an expectation.

  The decision-maker knows $\mathcal{D}$ but does not know the specific bandit problem $\nu = (\nu_a)_{a \in \{1 \ldots K\}} \in \mathcal{D}^K$ at hand.

- Strategy ψ consistent wrt. a model $\mathcal{D}$: when $\forall \nu \in \mathcal{D}^K$, $\forall \alpha \in (0,1]$, $\forall a$ s.t. $\Delta_a > 0$ $\mathbb{E}_\nu[N_a(T)] = o(T^\alpha)$

Lemma (Fundamental inequality for stochastic bandits): For all strategies ψ,

For all bandit problems $\nu = (\nu_a)_{a \in \{1 \ldots K\}}$ and $\nu' = (\nu'_a)_{a \in \{1 \ldots K\}}$ in $\mathcal{D}^K$ with $\nu_a \ll \nu'_a$ for all a,

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

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

$$\geq \mathrm{KL}\Big(\mathrm{Ber}\big(\mathbb{E}_\nu[Z]\big), \mathrm{Ber}\big(\mathbb{E}_{\nu'}[Z]\big)\Big)$$

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

Proof: • The inequality $\geq$ 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 $\mathbb{P}_{\underline{\nu}}^{H_T}$ 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, ie., that the $U_0, Y_1, U_1, \dots Y_T, U_T$ are the projections on each component,
- $\mathbb{P}_{\underline{\nu}}$ is given by

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

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

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

A similar construction can be done for the bandit problem $\underline{\nu}'$.

Now, the equality $\mathbb{P}_{\underline{\nu}}(Y_{t+1} \in B' \text{ and } U_{t+1} \in B \mid H_t) = \nu_{\psi_t(H_t)}(B)\ \eta(B)$

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

Let us check the assumptions of our chain rule:

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

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

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

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

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

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</u> giving the lower bound:

<u>Theorem:</u> 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 $\underline{\nu} = (\nu_a)_{a\in\{1..K\}} \in \mathcal{D}^K$,
For all suboptimal arms a (ie, arms a with $\Delta_a > 0$),

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

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>Correction for Exercise #3</u>: The proof of this theorem based on the lemma above was your Exercise #2.

See next page.

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, $\underline{\nu}$ and a s.t. $\Delta_a > 0$
- Fix an alternative model $\underline{\nu}'$ 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, $\underline{\nu}$ and $\underline{\nu}'$ only differ at a; a is the unique optimal arm in $\underline{\nu}'$

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

Our fundamental inequality yields, since $\underline{\nu}$ and $\underline{\nu}'$ only differ at a:

$$\mathbb{E}_{\underline{\nu}}[N_a(T)] \ KL(\nu_a, \nu'_a) \geqslant KL\Big(\mathrm{Ber}\big(\mathbb{E}_{\underline{\nu}}[N_a(T)/T]\big), \mathrm{Ber}\big(\mathbb{E}_{\underline{\nu}'}[\tfrac{N_a(T)}{T}]\big)\Big)$$

$$\geqslant -\ln 2 + \big(1 - \mathbb{E}_{\underline{\nu}}[N_a(T)/T]\big) \ln \frac{1}{1 - \mathbb{E}_{\underline{\nu}'}[N_a(T)/T]}$$

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

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

$$= \underbrace{p \ln \frac{1}{q}}_{\geqslant 0} + (1-p) \ln \frac{1}{1-q} + \underbrace{\big(p \ln p + (1-p)\ln(1-p)\big)}_{\geqslant -\ln 2}$$

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

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 $\underline{\nu}$, a is suboptimal: $\mathbb{E}_{\underline{\nu}}[N_a(T)/T] \longrightarrow 0$

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

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

↳ 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:

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

thus

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

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

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 $1/\mathcal{K}_{\inf}(\nu_a, \mu^*)$ lower bound.

<u>Remark:</u> Non-asymptotic lower bounds are possible (for « super-consistent » strategies and « well behaved » models), but are heavily technical.

<u>Exercise 1.</u> $1/\mathcal{K}_{\inf}(\nu_a, \mu^*, \mathcal{D})$ 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:

$$\forall \underline{\nu} \in \mathcal{P}([0,1])^K, \qquad \forall a \text{ s.t. } \Delta_a > 0,$$

$$\mathbb{E}_{\underline{\nu}}[N_a(T)] \leqslant \frac{8}{\Delta_a^2} \ln T + 2.$$

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 those $8/\Delta_a^2$ and $2/\Delta_a^2$ constants compare to $1/\mathcal{K}_{\inf}(\nu_a, \mu^*, \mathcal{P}([0,1]))$?

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

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

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

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

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

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

Even better, show the stronger form:

$$\sup_{\substack{Z \text{ } \mathcal{F}\text{-measurable} \\ \text{taking values in } [0,1]}} \left| \mathbb{E}_{\mathbb{P}}[Z] - \mathbb{E}_{\mathbb{Q}}[Z] \right| \leqslant \sqrt{\frac{1}{2} \mathrm{KL}(\mathbb{P}, \mathbb{Q})}$$

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

Lower bound on the regret for adversarial bandits

$\leftrightarrow$ Minimax lower bound on the regret for stochastic bandits

We discussed so far $\mathbb{E}_{\nu}[N_a(T)]$ in terms of distribution-dependent bounds like

$$C_a(\nu) \ln T$$

What about distribution-free / minimax bounds?

To that end we restrict our attention to the model $\mathcal{D} = \mathcal{P}([0,1])$, the set of all probability distributions over $[0,1]$.

Stochastic bandits

With each arm a is associated $\nu_a \in \mathcal{P}([0,1])$

For $t = 1, 2, \ldots$

- The decision maker picks $I_t \in \{1..K\}$
- Her reward Y_t, which is such that $Y_t \mid I_t \sim \nu_{I_t}$, is her only piece of information

Aim: control the regret

$$\bar{R}_T = T \max_{a=1..K} E(\nu_a) - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right]$$

Adversarial bandits

An opponent selects the payoffs $g_{j,t}$

For $t = 1, 2, \ldots$

- The opponent picks $(g_{1t} \ldots g_{Kt}) \in [0,1]^K$
- while, simultaneously, The decision-maker picks $I_t \in \{1..K\}$
- Her payoff is $g_{I_t,t}$ and this is the only piece of information she gets

Aim: control the regret

$$R_T = \max_{k=1..K} \sum_{t=1}^{T} g_{kt} - \sum_{t=1}^{T} g_{I_t,t}$$

Typical adversarial results (Auer, Cesa-Bianchi, Freund, Schapire, 2002, later improved by Audibert and Bubeck, 2009):

Strategies such that for all opponents picking gains in $[0,1]$, for all $T \geq 1$,

$$\begin{cases} \text{with probability at least } 1-\delta, & R_T \leq C\sqrt{TK\ln(K/\delta)} \\ \mathbb{E}[R_T] \leq C\sqrt{TK\ln K} & \end{cases}$$

where the probability and $\mathbb{E}$ are wrt decision-maker's internal randomization

for some numerical constant C

For "oblivious" opponents (ie, when the g_{kt} do not "react" to the decision-maker's actions): the $\sqrt{\ln K}$ can be dropped.

It is in particular the case when $g_{kt} \sim \nu_k$ $\forall t$ in an independent way

 Lecture 3 — February 22, 2017 Lecture 3 — February 22, 2017

In this stochastic model:

$$\mathbb{E}[R_T] = \mathbb{E}\Big[\max_{k=1..K} \sum_{t=1}^{T} g_{k,t}\Big] - \mathbb{E}\Big[\sum_{t=1}^{T} g_{I_t,t}\Big]$$

$g_{I_t,t}$ is Y_t

$$\geq T \max_{k=1..K} \mathbb{E}[g_{k,1}] - \mathbb{E}\Big[\sum_{t=1}^{T} Y_t\Big]$$

$$= T \max_{k=1..K} \mathbb{E}(\nu_k) - \mathbb{E}\Big[\sum_{t=1}^{T} Y_t\Big] = \overline{R}_T$$

The adversarial results entail in particular that there exists a strategy of the decision-maker such that

$$\sup_{\nu_1 \dots \nu_K \in \mathcal{P}([0,1])} \overline{R}_T \leq \sup_{\substack{\text{opponents} \\ \text{picking } g_{j,t} \in [0,1]}} \mathbb{E}[R_T] \leq C\sqrt{TK}$$

for some numerical constant C

<u>Lower bound:</u>

<u>Theorem</u>: For all (randomized) strategies of the decision-maker, for all $K \geq 2$ and $T \geq K \ln 2 / 2$,

$$\sup_{\substack{\text{opponents} \\ \text{picking } g_{j,t} \in [0,1]}} \mathbb{E}[R_T] \geq \sup_{\substack{\text{individual} \\ \text{sequences } g_{j,t} \in [0,1]}} \mathbb{E}[R_T] \geq \sup_{\nu_1, \dots \nu_K \in \mathcal{P}([0,1])} \overline{R}_T \geq \frac{1}{20}\sqrt{TK}$$

the only inequality we need to prove!

<u>Proof</u>: (i) Bandit problem $\nu^{(0)} = \big(\mathrm{Ber}(1/2), \dots \mathrm{Ber}(1/2)\big)$

versus bandit problems $\nu^{(i)} = \big(\mathrm{Ber}(1/2), \dots \mathrm{Ber}(1/2), \mathrm{Ber}(1/2+\varepsilon), \mathrm{Ber}(1/2), \dots\big)$

for $i \in \{1..K\}$ and $\varepsilon \in (0, 1/2)$ — in i-th position

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

The fundamental inequality indicates that

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

$$= \mathbb{E}_{\nu^{(0)}}[N_k(T)] \, \mathrm{KL}\big(\nu_k^{(0)}, \nu_k^{(k)}\big) \leq T/K \; \mathrm{kl}\big(1/2, 1/2+\varepsilon\big)$$

Using Pinsker's inequality:

$$\mathrm{kl}\Big(\mathbb{E}_{\nu^{(0)}}\Big[\frac{N_k(T)}{T}\Big], \mathbb{E}_{\nu^{(k)}}\Big[\frac{N_k(T)}{T}\Big]\Big) \geq 2\Big(\mathbb{E}_{\nu^{(0)}}\Big[\frac{N_k(T)}{T}\Big] - \mathbb{E}_{\nu^{(k)}}\Big[\frac{N_k(T)}{T}\Big]\Big)^2$$

Solving for $\mathbb{E}_{\nu^{(k)}}\left[\frac{N_k(T)}{T}\right]$:

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

$$\leq \frac{1}{K} + \sqrt{\frac{T}{2K}\, kl(1/2, 1/2+\varepsilon)}$$

$$\leq \frac{1}{K} + \sqrt{(2\ln 2)\frac{T\varepsilon^2}{K}} \qquad \text{for } \varepsilon \in \left(0, 1/2\sqrt{2}\right]$$

Indeed: $\varepsilon \in (0, 1/2\sqrt{2}]$ entails $4\varepsilon^2 \leq 1/2$, as $\ln\left(1/_{1-u}\right) \leq (2\ln 2)\, u \quad \forall u \in (0, 1/2]$

and $kl(1/2, 1/2 - \varepsilon) = \frac{1}{2}\ln\left(\frac{1}{1-2\varepsilon}\right) + \frac{1}{2}\ln\left(\frac{1}{1+2\varepsilon}\right) = \frac{1}{2}\ln\frac{1}{1-4\varepsilon^2}$

$$\leq (4\ln 2)\,\varepsilon^2$$

(2) Let's go back to the regret.

In $\nu^{(k)}$:
$$\overline{R}_T = \sum_{a \neq k} \underset{\substack{\uparrow \\ \text{gap} \\ \text{of arm } a}}{\varepsilon}\ \underset{\substack{\uparrow \\ \text{number of} \\ \text{times } a \text{ is pulled}}}{\mathbb{E}_{\nu^{(k)}}[N_a(T)]} = \varepsilon\left(T - \mathbb{E}_{\nu^{(k)}}[N_k(T)]\right)$$

$$\geq T\varepsilon\Big(\underbrace{1 - \frac{1}{K}}_{\geq 1/2} - \sqrt{(2\ln 2)\, T\varepsilon^2/K}\Big)$$

to be optimized over $\varepsilon \in (0, 1/2\sqrt{2}]$

Optimal ε s.t. $\frac{1}{2} - 2\varepsilon\sqrt{(2\ln 2)\, T/K} = 0$, ie, $\varepsilon = \frac{1}{4} \times 1/\sqrt{(2\ln 2)\, T/K}$

which is $\leq 1/2\sqrt{2}$ as soon as $T\ln 2/K \geq 1/2$,

For this ε, the lower bound is ie $T \geq K\ln 2/2$

$$T\varepsilon\Big(1/2 - \varepsilon\underbrace{\sqrt{(2\ln 2)\, T/K}}_{=1/4}\Big) = T\varepsilon/4 = \frac{1}{16\sqrt{2\ln 2}}\sqrt{TK} \geq \frac{1}{20}\sqrt{TK}$$

Note about the constants: By further restricting the considered T (ie, by imposing $T \geq \gamma K$ with a larger γ), we can improve the $\frac{1}{20}\sqrt{TK}$ bound (ie, get $c'\sqrt{TK}$ with $c' > 1/20$).

This is done in the upper bound on $\ln(1/_{1-u})$: eg, $\leq 4\ln(4/3)\, u$ for $u \in [0, 1/4]$

↳ Entails: $\sup \overline{R}_T \geq \frac{\sqrt{2}-1}{\sqrt{32\ln(4/3)}}\sqrt{TK} \geq 0{,}136\sqrt{TK}$ for $T \geq \frac{1}{4\ln(4/3)}K \simeq 0{,}87\, K$

Open question:

So, the minimax rates for the regret for stochastic bandits or oblivious opponents are $\sqrt{TK}$.

What is the minimax rate against general, reactive opponents?

→ Should the upper bound be improved?

→ Should the proof technique be improved?
(in particular, look for sequences of payoffs $g_{k,t}$ with real and strong correlations / dependencies in the past).

Stochastic bandits : What about arms indexed by a continuum?

Setting 1 : Arms indexed by $x \in A$, where A is some possibly large set
With each arm $x \in A$ is associated a probability distribution ν_x over $\mathbb{R}$ s.t. $E(\nu_x)$ exists
At each round, the decision-maker picks $I_t \in A$, 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 : $f : x \in A \longmapsto E(\nu_x)$ is the mean-payoff function

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

Setting 2 : [special case] $\longrightarrow$ Noisy optimization of a function.
We fix $f : A \to \mathbb{R}$
The noise is given by a sequence of iid random variables $\varepsilon_1, \varepsilon_2, \ldots$
When $I_t \in A$ 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 $\underline{\nu} = (\nu_x)_{x \in A}$
The regret can be controlled (in a non-uniform way) against $\mathcal{F}$ if :

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

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

Ex: $A = \{1, \dots K\}$ and $\underbrace{\mathcal{F} = \left(\mathcal{P}([0,1])\right)^K}_{\text{the case of finitely many arms with bounded distributions}}$, the set of all K-tuples of probability distributions over $[0,1]$

$\rightarrow$ UCB does the job.

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

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, $\mathcal{S} = \{x \in [0,1] : \exists t \mid \mathbb{P}\{I_t = x\} > 0 \text{ under } (\delta_0)_{x\in[0,1]}\}$ is countable. In particular, $[0,1] \setminus \mathcal{S}$ is non empty.

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

With probability 1, it thus never hits x_0.

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

Actually, continuity is necessary and sufficient for the regret to be controlled, as long as A is not too large.

Theorem: Let A be a metric space and let $\mathcal{F}^{\text{cont}}$ be the set of bandit problems $(\nu_x)_{x\in A}$

with: $\rightarrow$ $\forall x$, ν_x is a distribution over $[0,1]$

$\rightarrow$ a continuous mean-payoff function $f: x \mapsto E(\nu_x)$

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

<u>Corollary.</u> 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.

<u>Exercise 2.</u> 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:

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

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

Hint: First, prove a performance bound by splitting $[0,1]$ into $[(i-1)/K, i/K]$ with $i = 1, \ldots, K$ for a fixed K, where each bin $[(i-1)/K, i/K]$ 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.

<u>Proof of the Corollary:</u>

We endow $\mathcal{A}$ with the discrete topology, ie, 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 A is separable:** let $(x_n)_{n\in\mathbb{N}}$ be a collection of points in A that is dense.

In particular, the probability distribution $\lambda = \sum_{n\geq 0} \frac{1}{2^{n+1}} \delta_{x_n}$ is such that $\lambda(V) > 0$ for all open sets $V \subset A$.

We pick elements $J_1 \; J_2 \; I_1^{(2)} \dots J_t \; I_1^{(t)} \dots I_{t^2-1}^{(t)}$ as follows:

$$
\begin{array}{lll}
J_1 & & \leftarrow \text{regime } 1 \\
J_2 & I_1^{(2)} \dots I_3^{(2)} & \leftarrow \quad 2 \quad 2^2 \text{ rounds} \\
\vdots & \ddots & \quad\quad \vdots \\
J_t & I_1^{(t)} \dots I_{t^2-1}^{(t)} & \leftarrow \quad t \quad t^2 \text{ rounds}
\end{array}
$$

where $\begin{cases} \text{the } J_s \text{ are drawn at random according to } \lambda \\ \text{the } I_s^{(r)},\ 1 \leq s \leq r^2-1, \text{ follow from the UCB strategy with arms } J_1 \dots J_r \end{cases}$

regime r starts at time $S_{r-1}+1$ where $[1+2^2+\dots+(r-1)^2] = S_{r-1}$

In regime r:
$$r^2 \max_{s\leq r} \mu_{J_s} - \mathbb{E}\left[\sum_{s=1}^{r^2} Y_{S_{r-1}+s}\right]$$
$$\leq 1 + c\sqrt{r^3 \ln r}$$

1: for J_r

c: well-chosen numerical constant

$\sqrt{r^3\ln r}$: distribution-free regret bound for UCB on r^3-1 steps with r arms (we saw this bound as an exercise)

Let $\varepsilon > 0$, let $\tilde{r}_\varepsilon$ the first (random) time when $\mu_{J_r} = f(J_r) \geq \sup_{x\in A} f(x) - \varepsilon$

We have $\tilde{r}_\varepsilon < +\infty$ a.s. because:

- by continuity of f, there exists an open set V_ε with
$$\forall x' \in V_\varepsilon, \quad f(x') \geq \sup_A f - \varepsilon;$$

– we have $\tilde{r}_\varepsilon \leqslant \inf\{ r \geqslant 1 : \; J_r \in V_\varepsilon \} < +\infty$ a.s.

as this random variable follows a geometric distribution with parameter $d(V_\varepsilon) > 0$.

For $r \geqslant \tilde{r}_\varepsilon$, $\displaystyle \max_{s \leq r} \mu_{J_s} + \varepsilon \geqslant \sup_{\mathcal{A}} f$

So that

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

$$\leqslant \underbrace{\sum_{r=1}^{\tilde{r}_\varepsilon - 1} r^2}_{\substack{\text{lengths of regimes } r \leq \tilde{r}_\varepsilon - 1 \\ < +\infty \text{ a.s.}}} + T\varepsilon + \overbrace{\sum_{r=\tilde{r}_\varepsilon}^{r_T - 1} \left(1 + c\sqrt{r^3 \ln T} \right)}^{\text{the "complete regimes"}} + \underbrace{r_T^2}_{\text{regime } r_T \text{ may be incomplete}}$$

where r_T is such that time T belongs to regime r_T: we have r_T^3 of order T i.e. r_T of order $T^{1/3}$

and

$$\sum_{r \leq r_T} \left(1 + c r^{3/2} \sqrt{\ln r} \right)$$

$$\leqslant \sum_{r \leq r_T} \left(1 + c r^{3/2} \sqrt{\ln r_T} \right) = O\left(r_T^{5/2} \sqrt{\ln r_T} \right)$$

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

Thus,

$$\limsup_{T \to +\infty} \frac{\overline{R}_T}{T} \leqslant \varepsilon \quad \text{a.s.}$$

but since $\overline{R}_T$ is a deterministic quantity and this is true $\forall \varepsilon > 0$, we have

$$\lim_{T \to +\infty} \frac{\overline{R}_T}{T} = 0 \quad \text{as requested.}$$

 Lecture 3 — February 22, 2017 Lecture 3 — February 22, 2017

2) <u>If $\mathcal{A}$ is not separable:</u>

* 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 $f^{(a)} : x \in \mathcal{A} \longmapsto \left(1 - \dfrac{d(x,a)}{\rho/2}\right)^+$; 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 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 $\mathcal{L}_t$ be the law of I_t under $(\delta_0)_{x\in\mathcal{A}}$ and let $\mathcal{L} = \sum_{t\geq 1} \frac{1}{2^t} \mathcal{L}_t$.

There exists $a \in \mathcal{A}$ s.t. $\mathcal{L}\big(B(a,\rho/2)\big) = 0$, that is, s.t.,
$$\forall t \geq 1, \quad \mathbb{P}\big(I_t \in B(a,\rho/2) \text{ under } (\delta_0)_{x\in\mathcal{A}} \big) = 0.$$

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 <u>not</u> controlled against $\nu^{(a)} \in \mathcal{F}^{\mathrm{cont}}$.