Lecture #7: $\sqrt{KT}$ distribution free bound and bandits with a continuum of arms

We have shown (in exercise session #4) · a minimax lower bound of order $\sqrt{KT}$ for stochastic bandits

· distribution free upper bounds of order $\sqrt{KT \ln T}$ for UCB and SE.

Can we get a $\sqrt{KT}$ upper bound?

MOSS algorithm

(Minimax Optimal Strategy in the Stochastic case of bandit problems)

Index policy relying on $U_k(t) = \hat{\mu}_k(t) + \sqrt{\frac{1}{2N_k(t)} \ln_+\left(\frac{t}{KN_k(t)}\right)}$

where $\ln_+ = \max(\ln, 0)$.

ie algo is defined as

For $t = 1, \dots, K$: pull $a_t = t$

For $t \geq K+1$: pull $a_t \in \underset{k \in [K]}{\operatorname{argmax}} \; U_k(t-1)$

Difference with UCB:

bonus

$$\sqrt{\frac{2 \ln t}{N_k(t)}} \quad \text{vs} \quad \sqrt{\frac{\ln_+(t/KN_k(t))}{2N_k(t)}}$$

$\longrightarrow$ no exploration after k was pulled $\frac{t}{K}$ times (still exploration)

Theorem

MOSS satisfies for bandit model $\mathcal{D} = \mathcal{P}([0,1])$

$$\sup_{\nu \in \mathcal{D}^K} R_T(\text{MOSS}, \nu) \leq K-1 + 45\sqrt{KT}$$

$\to$ minimax optimal, up to constant factor (the 45 constant can still be improved)

Proof: First step for $t \geq K+1$, $U_{k^*}(t-1) \leq U_{a_t}(t-1)$ by defn of algorithm

Thus
$$R_T \leq \underbrace{K-1}_{\substack{\text{at most } K-1 \\ \text{suboptimal pulls in} \\ \text{first } K \text{ steps}}} + \sum_{t=K+1}^{T} \mathbb{E}[\mu^* - U_{k^*}(t-1)] + \underbrace{\sum_{t=K+1}^{T} \mathbb{E}[U_{a_t}(t-1) - \mu_{a_t}]}_{\leq \sqrt{KT} + \sum_{t=K+1}^{T} \mathbb{E}\left[\left(U_{a_t}(t-1) - \mu_{a_t} - \sqrt{\frac{K}{t}}\right)_+\right]}$$

Second step: control of each $\mathbb{E}[\mu^* - U_{k^*}(t)]$ term by $20\sqrt{\frac{K}{t}}$ (for $t \geq K$)

for that:
$$\mathbb{E}[\mu^* - U_{k^*}(t)] \leq \mathbb{E}[(\mu^* - U_{k^*}(t))_+]$$

$$\leq \sum_{l=0}^{+\infty} \mathbb{E}\left[(\mu^* - U_{k^*}(t))_+ \mathbb{1}_{\{N_{k^*}(t) \in [x_{l+1}, x_l[\}}\right]$$

where $x_l = \beta^{-l} \frac{t}{K}$ for some fixed $\beta > 1$

$$+ \mathbb{E}\left[(\mu^* - U_{k^*}(t))_+ \mathbb{1}_{\{N_{k^*}(t) \geq x_0\}}\right]$$

Now, $U_{k^*}(t) = \hat{\mu}_{k^*}(t) + \begin{cases} 0 & \text{if } N_{k^*}(t) \geq \frac{t}{K} = x_0 \\ \sqrt{\frac{1}{2N_{k^*}(t)} \ln\left(\frac{t}{KN_{k^*}(t)}\right)} \geq \underbrace{\sqrt{\frac{1}{2x_l} \ln\left(\frac{t}{Kx_l}\right)}}_{:= \varepsilon_l} & \text{if } N_{k^*}(t) \in [x_{l+1}, x_l] \end{cases}$

So $$\mathbb{E}[\mu^* - U_{k^*}(t)] \leq \mathbb{E}\left[(\mu^* - \hat{\mu}_{k^*}(t))_+ \mathbb{1}_{\{N_{k^*}(t) \geq \frac{t}{K}\}}\right] + \sum_{l=0}^{+\infty} \mathbb{E}\left[(\mu^* - \hat{\mu}_{k^*}(t) - \varepsilon_l)_+ \mathbb{1}_{\{N_{k^*}(t) \in [x_{l+1}, x_l]\}}\right]$$

Lemma: $$\mathbb{E}\left[(\mu^* - \hat{\mu}_{k^*}(t) - \varepsilon)_+ \mathbb{1}_{\{N_{k^*}(t) \geq n_0\}}\right] \leq \frac{1}{\sqrt{n_0}} e^{-2n_0\varepsilon^2}$$

Proof of the lemma:

$$\mathbb{E}\left[(\mu^* - \hat{\mu}_{k^*}(t) - \varepsilon)_+ \mathbb{1}_{\{N_{k^*}(t) \geq n_0\}}\right] = \int_0^{+\infty} \mathbb{P}\left(\mu^* - \hat{\mu}_{k^*}(t) - \varepsilon \geq u \text{ and } N_{k^*}(t) \geq n_0\right) du$$

$$= \int_0^{+\infty} \mathbb{P}\left(Z_t^* \geq (\varepsilon + u) N_{k^*}(t) \text{ and } N_{k^*}(t) \geq n_0\right) du$$

where $Z_t^* = N_{k^*}(t)(\mu^* - \hat{\mu}_{k^*}(t)) = \sum_{s=1}^{t} (\mu^* - X_{k^*}(s)) \mathbb{1}_{\{A_s = k^*\}}$ is a martingale

see proof of UCB regret

and for all $x \in \mathbb{R}$, $S_{x,t} = e^{x Z_t - \frac{x^2}{8} N_{k^*}(t)}$ is a supermartingale

Thus by Markov-Chernoff, we continue the bounding as, for $x > 0$

$$= \int_0^{+\infty} \mathbb{P}\left\{ e^{x Z_t^* - \frac{x^2}{8} N_{k^*}(t)} \geq \exp\left(N_{k^*}(t)\left(x(\varepsilon+u) - \frac{x^2}{8}\right)\right) \text{ and } N_{k^*}(t) \geq n_0 \right\} du$$

$x = 4(\varepsilon + u)$ so that $x(\varepsilon+u) - \frac{x^2}{8} = 2(\varepsilon+u)^2$

$$\leq \int_0^{+\infty} \sum_{l=n_0}^{+\infty} e^{-2l(\varepsilon+u)^2} \, \mathbb{E}\left[S_{4(\varepsilon+u),t} \mathbb{1}_{\{N_{k^*}(t) = l\}}\right] du$$

$$\leq \int_0^{+\infty} e^{-2n_0(\varepsilon^2+u^2)} \mathbb{E}\left[S_{4(\varepsilon+u),r} \mathbb{1}_{\{N_{a^*}(r)\geq n_0\}}\right] du$$

we know $\mathbb{E}[S_{4(\varepsilon+u),r}] \leq 1$ (cond $s \geq d$)

So all in all:

$$\mathbb{E}\left[(\mu^* - \hat{\mu}_{a^*}(r) - \varepsilon)_+ \mathbb{1}_{\{N_{a^*}(r)\geq n_0\}}\right] \leq e^{-2n_0\varepsilon^2} \underbrace{\int_0^{+\infty} e^{-2n_0 u^2} du}_{\text{integral of a Gaussian density (up to renorm)}} = e^{-2n_0\varepsilon^2} \cdot \sqrt{\frac{\pi}{8n_0}} \leq \frac{e^{-2n_0\varepsilon^2}}{\sqrt{n_0}}$$

Going back to the main proof:

$$\mathbb{E}[\mu^* - U_{a^*}(r)] \leq \sqrt{\frac{K}{r}} + \sum_{\ell=0}^{+\infty} \underbrace{\frac{1}{\sqrt{x_{\ell+1}}} e^{-2x_{\ell+1}\varepsilon_\ell^2}}$$

$$\frac{1}{\sqrt{x_{\ell+1}}} \exp\left(-2x_{\ell+1} \cdot \frac{1}{2x_\ell} \ln\left(\frac{r}{Kx_\ell}\right)\right)$$

$$= \frac{1}{\sqrt{x_{\ell+1}}} \exp\left(-\frac{1}{\beta} \cdot \ell \ln(\beta)\right)$$

$$= \sqrt{\frac{K}{r}} \beta^{\frac{\ell+1}{2}} \exp\left(-\frac{\ell}{\beta} \ln(\beta)\right)$$

$$= \sqrt{\frac{K}{r}} \beta^{\frac{1}{2} + \ell\left(\frac{1}{2} - \frac{1}{\beta}\right)}$$ we want $\beta \in (1,2)$

Taking eg. $\beta = \frac{3}{2}$: $\mathbb{E}(\mu^* - U_{a^*}(r)) \leq \sqrt{\frac{K}{r}} + \sqrt{\frac{K}{r}} \beta^{1/2} \cdot \sum_{\ell=0}^{+\infty} \left(\beta^{\left(\frac{1}{2}-\frac{1}{\beta}\right)}\right)^\ell$

$$= \sqrt{\frac{K}{r}} \left(1 + \underbrace{\sqrt{\frac{3}{2}} \cdot \frac{1}{1-\alpha}}_{\leq 19}\right)$$ with $\alpha = \left(\frac{3}{2}\right)^{\left(\frac{1}{2}-\frac{2}{3}\right)} \in (0,1)$

$$\leq 20\sqrt{\frac{K}{r}}.$$

third step: $\displaystyle\sum_{t=K+1}^{T} \mathbb{E}\left[\left(U_{a_t}(t-1) - \mu_{a_t} - \sqrt{\tfrac{K}{T}}\right)_+\right]$ is $\leq 4\sqrt{KT}$

$$= \sum_{r=K}^{T-1} \mathbb{E}\left[\left(U_{a_{r+1}}(r) - \mu_{a_{r+1}} - \sqrt{\tfrac{K}{T}}\right)_+\right]$$

$$\sum_{r=K}^{T-1} \mathbb{E}\left[\left(U_{a_{r+1}}(r) - \mu_{a_{r+1}} - \sqrt{\tfrac{K}{T}}\right)_+\right] = \sum_{k=1}^{K}\sum_{l=1}^{T}\sum_{r=K}^{T-1} \mathbb{E}\left[\left(U_k(r) - \mu_k - \sqrt{\tfrac{K}{T}}\right)_+ \mathbb{1}_{\{a_{r+1}=k\}}\, \mathbb{1}_{\{N_k(r)=l\}}\right]$$

we now use $\displaystyle\left(U_k(r) - \mu_k - \sqrt{\tfrac{K}{T}}\right)_+ \leq \left(\hat{\mu}_k(t) - \mu_k - \sqrt{\tfrac{K}{T}}\right)_+ + \begin{cases} 0 & \text{if } N_k(r) \geq \frac{T}{K} \\ \sqrt{\frac{1}{2N_k(r)}\ln\left(\frac{T}{KN_k(r)}\right)} & \text{if } N_k(r) < \frac{T}{K} \end{cases}$

also smaller than $\sqrt{\frac{1}{2N_k(r)}\ln\left(\frac{T}{N_k(r)}\right)}$

and get therefore the upper bound:

$$\leq \sum_{k=1}^{K}\sum_{l=1}^{T}\sum_{r=K}^{T-1} \mathbb{E}\left[\left(\hat{\mu}_k(r) - \mu_k - \sqrt{\tfrac{K}{T}}\right)_+ \mathbb{1}_{\{a_{r+1}=k\}}\, \mathbb{1}_{\{N_k(r)=l\}}\right] + \sum_{k=1}^{K}\sum_{l=1}^{\lfloor T/K\rfloor} \sqrt{\frac{1}{2l}\ln\left(\frac{T}{Kl}\right)}\; \underbrace{\mathbb{E}\left[\sum_{r=K}^{T-1} \mathbb{1}_{\{N_k(r)=l\}}\, \mathbb{1}_{\{a_{r+1}=k\}}\right]}_{\leq 1 \text{ a.s.}}$$

Also
$$\sum_{l=1}^{\lfloor T/K\rfloor} \sqrt{\frac{1}{2l}\ln\left(\frac{T}{Kl}\right)} \leq \int_0^{T/K} \sqrt{\frac{1}{2x}\ln\left(\frac{T}{Kx}\right)}\,dx$$

c.v. $u = \frac{T}{Kx}$ $\quad du = -\frac{T}{Kx^2}\,dx$ $\quad \frac{dx}{\sqrt{x}} = -\frac{K}{T}x^{3/2}\,du = -\sqrt{\frac{T}{K}}\,u^{-3/2}\,du$

$$\leq \sqrt{\frac{T}{2K}} \int_1^{+\infty} u^{-3/2}\sqrt{\ln(u)}\,du$$

cv $u = e^{v^2}$ $\quad du = 2v\,e^{v^2}\,dv$ $\quad u^{-3/2}\,du = 2v\,e^{-\frac{v^2}{2}}\,dv$ $\quad \sqrt{\ln u} = v$

$$= \sqrt{\frac{T}{2K}} \int_0^{+\infty} \underbrace{2v^2 e^{-\frac{v^2}{2}}\,dv}_{\text{variance of standard Gaussian, up to } \sqrt{2\pi} \text{ renormalization}}$$

$$= \sqrt{\pi}\sqrt{\frac{T}{K}}.$$

summarizing, we showed so far (in third step):

we will show that $\leq \sqrt{\frac{\pi}{2}}\sqrt{\frac{T}{K}}$ for each k

$$\sum_{r=K}^{T-1} \mathbb{E}\left[\left(U_{a_{r+1}}(r) - \mu_{a_{r+1}} - \sqrt{\tfrac{K}{T}}\right)_+\right] \leq \overbrace{\sum_{k=1}^{K}\sum_{l=1}^{T}\sum_{r=K}^{T-1} \mathbb{E}\left[\left(\hat{\mu}_k(r) - \mu_k - \sqrt{\tfrac{K}{T}}\right)_+ \mathbb{1}_{\{a_{r+1}=k\}}\, \mathbb{1}_{\{N_k(r)=l\}}\right]}$$
$$+ K\cdot\sqrt{\pi}\sqrt{T/K}$$

we resort again to $Z_{k,t} = N_k(\hat{\mu}_k(t) - \mu_k)$ martingale

$$S^{(k)}_{\lambda,t} = e^{\lambda Z_{k,t} - \frac{\lambda^2}{8} N_k(t)}$$ supermartingale

where $\lambda = 4\left(\sqrt{\frac{K}{T}} + u\right)$

For each k,

$$\sum_{\ell=1}^{T} \sum_{t=K}^{T-1} \mathbb{E}\left[\left(\hat{\mu}_k(t) - \mu_k - \sqrt{\frac{K}{T}}\right)_+ \mathbb{1}_{\{a_{t+1}=k\}} \mathbb{1}_{\{N_k(t)=\ell\}}\right] =$$

$$\sum_{\ell=1}^{T} \sum_{t=K}^{T-1} \int_0^{+\infty} \mathbb{P}\left(\lambda Z_{k,t} - \frac{\lambda^2}{8} N_k(t) \geq N_k(t)\left(\lambda\left(u + \sqrt{\frac{K}{T}}\right) - \frac{\lambda^2}{8}\right) \text{ and } a_{t+1}=k \text{ and } N_k(t)=\ell\right) du$$

$$\leq \sum_{\ell=1}^{T} \sum_{t=K}^{T-1} \int_0^{+\infty} e^{-2\ell\left(u^2 + \frac{K}{T}\right)} \mathbb{E}\left[\underbrace{S^{(k)}_{\lambda,t}}_{t} \underbrace{\mathbb{1}_{\{a_{t+1}=k\}} \mathbb{1}_{\{N_k(t)=\ell\}}}_{\text{the sum over } t \text{ of these will be } \leq 1}\right]$$

issue: depends on t...

but can be replaced in some sense, by $S^{(a)}_{\lambda,0} = 1$.

$$\leq \sum_{\ell=1}^{T} \int_0^{+\infty} e^{-2\ell\left(u^2 + \frac{K}{T}\right)} \mathbb{E}\left[\sum_{t=K}^{T-1} S^{(a)}_{\lambda,t} \mathbb{1}_{\{a_{t+1}=k\}} \mathbb{1}_{\{N_k(t)=\ell\}}\right] du$$

$$\leq \sum_{\ell=1}^{T} \int_0^{+\infty} e^{-2\ell\left(u^2 + \frac{K}{T}\right)} \mathbb{E}\left[S^{(k)}_{\lambda,\tau_\ell}\right] du$$

where $\tau_\ell = \inf\left\{t \in [T] : \begin{array}{l} a_{t+1} = k \text{ and} \\ N_k(t) = \ell \end{array}\right\} \wedge T$

$S_{x,r}^{(\ell)}$ supermartingale

τ is bounded $\longrightarrow$ we can apply optional stopping theorem ("théorème d'arrêt de Doob")

so that

$$\mathbb{E}\left[S^{(\ell)}_{x,\tau_\ell}\right] \leqslant \mathbb{E}\left[S^{(\ell)}_{x,0}\right] = 1.$$

So: $$\sum_{\ell=1}^{T}\sum_{r=K}^{T-1} \mathbb{E}\left[\left(\hat{\mu}_k(r) - \mu_k - \sqrt{\frac{K}{T}}\right)_+ \mathbb{1}_{\{a_{r+1}=k\}} \mathbb{1}_{\{N_k(r)=\ell\}}\right] \leqslant \sum_{\ell=1}^{T} \int_0^{+\infty} e^{-2\ell\left(u^2+\frac{K}{T}\right)} du$$

$$\leqslant \sum_{\ell=1}^{\infty} \frac{1}{\sqrt{\ell}} e^{-2\ell\frac{K}{T}}$$

) similar to the intermediate lemma in second step

$$\leqslant \int_0^{\infty} \frac{1}{\sqrt{x}} e^{-2x\frac{K}{T}} dx = \sqrt{\frac{T}{2K}} \int_0^{+\infty} \frac{e^{-u}}{\sqrt{u}} du$$

$u = v^2$

$\frac{du}{\sqrt{u}} = 2dv$

$$= \sqrt{\frac{T}{2K}} \cdot 2\int_0^{+\infty} e^{-v^2} dv = \sqrt{\frac{\pi}{2}}\sqrt{\frac{T}{K}}$$

Summarizing:

$$\sum_{r=K}^{T-1} \mathbb{E}\left[\left(U_{a_{r+1}}(r) - \mu_{a_{r+1}} - \sqrt{\frac{K}{T}}\right)_+\right] \leqslant \underbrace{\sum_{k=1}^{K}\sum_{\ell=1}^{T}\sum_{r=K}^{T-1} \mathbb{E}\left[\left(\hat{\mu}_k(r) - \mu_k - \sqrt{\frac{K}{T}}\right)_+ \mathbb{1}_{\{a_{r+1}=k\}} \mathbb{1}_{\{N_k(r)=\ell\}}\right]}_{K\sqrt{\frac{\pi}{2}}\sqrt{\frac{T}{K}}} + \underbrace{\sum_{k=1}^{K}\sum_{\ell=1}^{\lfloor T/K \rfloor} \sqrt{\frac{1}{2\ell}\ln\left(\frac{T}{K\ell}\right)}\, \mathbb{E}\left[\sum_{r=K}^{T-1} \mathbb{1}_{\{N_k(r)=\ell\}} \mathbb{1}_{\{a_{r+1}=k\}}\right]}_{+\sqrt{\pi}\sqrt{KT}}$$

$$\leqslant \sqrt{KT}\,.\, \sqrt{\pi}\left(1 + \frac{1}{\sqrt{2}}\right) \leqslant 4\sqrt{KT}.$$

General conclusion

Summarizing all steps, we bound the regret by

$$K-1 + \left(\sum_{t=K+1}^{T} 20\sqrt{\frac{K}{t-1}}\right) + \sqrt{KT} + 4\sqrt{KT} \leq K-1+5\sqrt{KT} + 20\int_0^T \sqrt{\frac{K}{s}}\,ds$$

$$= K-1+45\sqrt{KT} \quad \blacksquare$$

Bandits with continuum of arms

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 $a_t \in A$, gets a reward Y_t drawn at random according to ν_{a_t} (given a_t); and this is the only feedback she gets.

Definition $f: x \in A \longmapsto E(\nu_x)$ is the mean-payoff function.

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

Setting 2 (special case) $\Rightarrow$ noisy optimisation 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 $a_t \in A$ is picked, $Y_t = f(a_t) + \varepsilon_t$.

↳ 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 minimised

We always say that $(A, \mathcal{F})$ is tractable

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

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

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

Ex: $A = \{1, \dots, K\}$ and $\mathcal{F} = \mathcal{P}([0,1])^K$ $\rightarrow$ UCB does the job.

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

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

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

Since probability distributions can only have at most countably many atoms,

$$\mathcal{S} = \{ x \in [0,1] : \exists t \mid \mathbb{P}(a_t = x) > 0 \text{ under } (\delta_0)_{x \in [0,1]} \}$$

is countable. In particular,

we can consider $x_0 \in [0,1] \setminus \mathcal{S}$. The strategy then behaves the same under the problem $(\nu'_x)_{x \in [0,1]}$ in which

$$\begin{cases} \nu'_x = \delta_0 & \forall x \neq x_0 \\ \nu'_{x_0} = \delta_1 & \end{cases}$$

With proba 1, the strategy never pulls x_0.

Therefore, $Y_t = 0$ a.s. for any t and $R_T = T$.

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

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

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

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

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.
(see exercise session #5)

Proof of the corollary: we endow A with the discrete topology, i.e., choose the distance $d(x,y) = \mathbb{1}_{\{x \neq y\}}$. Then

1. All applications $f : A \to \mathbb{R}$ are continuous
2. A is separable if and only if 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.
We pick actions in a triangular fashion:

Regime 1 : UCB based on $x_1\, x_2$ (fresh start) : $a_1^{(1)}, \dots, a_4^{(1)}$

Regime r : UCB based on $x_1 \dots x_r\, x_{r+1}$ (fresh start) : $a_1^{(r)} \dots a_{(r+1)^2}^{(r)}$

In regime r:

starts at time $S_r + 1 = 2^2 + 3^2 + \dots + r^2 + 1$

(*) $$(r+1)^2 \max_{s \leq r} f(x_s) - \mathbb{E}\Big[\sum_{t=S_r+1}^{S_r+(r+1)^2} Y_t\Big] \leq \underbrace{c\sqrt{r^3 \ln r}}_{\substack{\text{distribution free bound}\\ \text{of UCB on } (r+1)^2 \text{ steps with}\\ (r+1) \text{ arms (see exercise session \#2)}}}$$

Now, let $\varepsilon > 0$ and let $\tilde{r}_\varepsilon \in \mathbb{N}^*$ s.t. $f(x_{\tilde{r}_\varepsilon}) \geq \sup_A f - \varepsilon$

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

In particular, $\max_{s \leq \tilde{r}_\varepsilon} f(x_s) \geq \sup_A f - \varepsilon$ (**)

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

so r_T is of the order of $T^{1/3}$, i.e. $r_T = O(T^{1/3})$.

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

$$R_T = T \sup_A f - \mathbb{E}\Big[\sum_{t=1}^{T} Y_t\Big] = \text{sum of the regrets of each regime}$$

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

$$\leq T\varepsilon + \sum_{r=\tilde{r}_\varepsilon}^{r_T - 1} r^{3/2} \sqrt{\ln r_T}$$

$$\leq T\varepsilon + O\big(r_T^{5/2} \sqrt{\ln r_T}\big)$$

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

All in all, $\limsup_T \frac{R_T}{T} \leq \varepsilon$ which is true for any $\varepsilon > 0$

that is $\lim \frac{R_T}{T} = 0$

2) If A is not separable

- We use the following characterisation of separability (which relies on Zorn's lemma).

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

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

$\Rightarrow$ No probability distribution over A can give a positive mass to all these balls.

- we consider the bandit models $\nu^{(a)}$ inducing mean-payoff function

$$f^{(a)} : x \in A \longmapsto \left(1 - \frac{d(x,a)}{\rho/2}\right)_+$$

continuous

in particular, $\nu_x^{(a)} = \delta_0$ for $x \notin B(a, \rho/2)$

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

As only countably many balls can have a positive mass under λ, there exists a s.t. $\lambda(B(a,\rho/2)) = 0$, that is s.t.

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

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