Lecture #9: Pure exploration

All previous lectures: maximise cumulative reward

→ exploration/exploitation trade-off.

In some applications, there is no price for exploring.

Think for example of a researcher testing drugs on mice / artificial human cells or testing products on some people before commercialisation.

Share similarities with regret minimisation, but good algorithms are actually different.

Setting (simple regret)

At each round $t = 1, \dots, T$:

- pulls an arm $a_t \in [K]$
- observes $X_{a_t}(t) \sim \nu_{a_t} \in \mathcal{D}$

We explore for T rounds and commit to best action at time $T+1$.

Goal: minimise simple regret

$$R_T^{\text{simple}} = \mathbb{E}\left[\mu^* - \mu_{a_{T+1}}\right].$$

Algorithm: Uniform exploration

For $t=1,\dots,T$:

choose $a_t = 1 + (t \bmod K)$

$A_{T+1} \in \arg\max_k \hat{\mu}_k(T)$.

Theorem: Uniform-Exploration satisfies for any $\nu \in \mathcal{P}([0,1])^K$

$$R_T^{simple} \leq \sum_{k, \Delta_k > 0} \Delta_k \exp\left(-2 \lfloor \tfrac{T}{K} \rfloor \Delta_k^2\right)$$

Proof Let k such that $\Delta_k > 0$.

$$\mathbb{P}\left(\hat{\mu}_{k^*}(T) \leq \hat{\mu}_k(T)\right) = \mathbb{P}\left(\hat{\mu}_{k^*}(T) - \hat{\mu}_k(T) \leq 0\right)$$

$N_{k^*}(T)$ and $N_k(T)$ are not random. We can directly apply Hoeffding inequality

$\geq \lfloor \frac{T}{K} \rfloor$

$$\mathbb{P}\left(\hat{\mu}_{k^*}(T) \leq \hat{\mu}_k(T)\right) \leq \exp\left(-2(N_k(T) + N_{k^*}(T))\Delta_k^2\right)$$

$$\leq \exp\left(-2 \lfloor \tfrac{T}{K} \rfloor \Delta_k^2\right).$$

$$R_T^{simple} = \sum_{k, \Delta_k > 0} \Delta_k \mathbb{P}(a_{T+1} = k)$$

$$\leq \sum_{k, \Delta_k > 0} \Delta_k \mathbb{P}\left(\hat{\mu}_{k^*}(T) \leq \hat{\mu}_k(T)\right)$$

$$\leq \sum_{k, \Delta_k > 0} \Delta_k \exp\left(-2 \lfloor \tfrac{T}{K} \rfloor \Delta_k^2\right) \quad \square$$

Theorem UE, distribution free bound

for any $\nu \in \mathcal{P}([0,1])^K$, $R_T^{simple} \leqslant 2\sqrt{\frac{K(\ln(K)\vee 1)}{T}}$

Proof

Actually, we could have written for any $\tilde{\Delta} \geqslant 0$

$$R_T^{simple} \leqslant \tilde{\Delta} + \sum_{k, \Delta_k > \tilde{\Delta}} \Delta_k \, \mathbb{P}(a_{T+1} = k)$$

$$\leqslant \tilde{\Delta} + K\tilde{\Delta} \exp\left(-2 \lfloor \tfrac{T}{K} \rfloor \tilde{\Delta}^2\right)$$

for any $\tilde{\Delta} \geqslant 0$.

Taking $\tilde{\Delta} = \sqrt{\frac{\ln(K)\vee 1}{2\lfloor T/K \rfloor}}$, we have:

$$R_T^{simple} \leqslant \tilde{\Delta} + \tilde{\Delta} K e^{-(\ln K \vee 1)} \leqslant 2\tilde{\Delta}$$

$$\leqslant \sqrt{2 \frac{\log K \vee 1}{\lfloor \frac{T}{K} \rfloor}}$$

if $T \leqslant 2K$, $R_T^{simple} \leqslant 1$ and the bound holds.

if $T \geqslant 2K$, $\lfloor \frac{T}{K} \rfloor \geqslant \frac{T}{2K}$

$$R_T^{simple} \leqslant 2\sqrt{\frac{K(\ln(K)\vee 1)}{T}}$$

$\square$

(see exercise session #6)

We can show that the minimax lower bound is larger than $c\sqrt{\frac{K}{T}}$, so Uniform-Exploration is nearly optimal in minimax sense.

Can we do better than UE? i.e. get rid of $\sqrt{\ln K}$ term.

Reduction from cumulative regret

For any strategy $(\pi_t)_{t=1,\dots,T}$; we can define $\tilde{\pi}$ s.t.

$$\tilde{\pi}_t = \pi_t \quad \text{for } t = 1, \dots, T$$

$$\mathbb{P}_{\tilde{\pi}}(a_{T+1} = k \mid \mathcal{F}_T) = \frac{N_k(T)}{T}.$$

Proposition for any instance ν, $R_T^{\text{simple}}(\nu, \tilde{\pi}) = \frac{R_T(\nu, \tilde{\pi})}{T}$.

Proof

$$R_T^{\text{simple}}(\nu, \tilde{\pi}) = \sum_{k=1}^{K} \mathbb{P}_{\tilde{\pi}}(a_{T+1} = k) \Delta_k$$

$$= \sum_{k=1}^{K} \mathbb{E}\left[\mathbb{P}_{\tilde{\pi}}(a_{T+1} = k \mid \mathcal{F}_T)\right] \Delta_k = \frac{1}{T} \sum_{k=1}^{K} \Delta_k \mathbb{E}[N_k(T)] = \frac{R_T(\nu, \pi)}{T} \quad \blacksquare$$

Corollary: There exists a strategy with a simple regret $R_T^{\text{simple}} \leq c\sqrt{\frac{K}{T}}$ for a universal constant c. (Reducing MOSS with $\tilde{\pi}$)

Best arm identification

Setting 1: (fixed confidence)

At each round $t = 1, \dots, \infty$:

- agent picks an arm $a_t \in [K]$ (based on previous observations)
- observes $X_{a_t}(t) \sim \nu_{a_t} \in \mathcal{D}$
- decides whether to continue sampling or stop.

If stop: return a final choice $\psi \in [K]$

The (random) stopping time is called τ

new

Goal: 1) Have a **sound** strategy with confidence level $\delta \in (0,1)$: $\mathbb{P}(\tau < \infty \text{ and } \mu_\psi < \mu^*) \leq \delta$ (confidence level) (for any ν)

2) Minimize the exploration time $\mathbb{E}[\tau]$.

Our algorithm will be built on the following lower bound.

Theorem (lower bound)

Let (π, τ, ψ) be a sound strategy for the bandit model $\mathcal{D}$, with confidence level $\delta \in (0,1)$, and let $\nu \in \mathcal{D}^K$. Then:

$$\mathbb{E}[\tau] \geq c^*(\nu) \ln\left(\frac{1}{4\delta}\right) \quad \text{where}$$

$$c^*(\nu)^{-1} = \sup_{\alpha \in \mathcal{P}_K} \left(\inf_{\nu' \in \mathcal{D}_{alt}(\nu)} \sum_{k=1}^{K} \alpha_k \, KL(\nu_k, \nu'_k) \right)$$

where $\mathcal{D}_{alt}(\nu) = \left\{ \nu' \in \mathcal{D}^K \;\middle|\; \operatorname{argmax}_k \mathbb{E}(\nu_k) \cap \operatorname{argmax}_k \mathbb{E}[\nu'_k] = \emptyset \right\}$ i.e. no arm is optimal for both ν and ν'

Another use of fundamental inequality (with stopping time)

Lemma: (admitted)

For all bandit problems $\nu = (\nu_k)_{k\in[K]}$ and $\nu' = (\nu'_k)_{k\in[K]}$ in $\mathcal{D}^K$ with $\nu_k \ll \nu'_k$ for all k,

for all strategies π, for any stopping time τ with respect to the filtration $(\sigma(H_t))$

and any random variable Z taking values in $[0,1]$, $\sigma(H_\tau)$-measurable,

$\sigma(H_\tau) = \{ A \in \sigma(H_\infty) \mid A \cap \{\tau \leq t\} \in \sigma(H_t) \text{ for all } t \}$

$$\sum_{k=1}^{K} \mathbb{E}_\nu[N_k(\tau)] \, KL(\nu_k, \nu'_k) \geq KL\big(Bn(\mathbb{E}_\nu[Z]), Bn(\mathbb{E}_{\nu'}[Z]) \big)$$

Proof of the Theorem

Assume $\mathbb{E}[\tau] < \infty$ (otherwise the result holds), so that $\mathbb{P}(\tau = \infty) = 0$.

Let $\nu' \in \mathcal{D}_{alt}(\nu)$. We define the $\sigma(H_\tau)$-measurable r.v.

$Z = \mathbb{1}_{\{\tau < \infty \text{ and } \psi \notin \operatorname{argmax}_k \mathbb{E}(\nu'_k)\}}$. Then the fundamental inequality (with stopping time)

yields:

$$\sum_{k=1}^{K} \mathbb{E}_\nu[N_k(\tau)] \, KL(\nu_k, \nu'_k) \geq KL\big(Bn(\underbrace{\mathbb{E}_\nu[Z]}_{\geq 1-\delta}), Bn(\underbrace{\mathbb{E}_{\nu'}[Z]}_{\leq \delta}) \big)$$

as π is sound with confidence level δ

$$\geq (1-\delta) \ln\left(\frac{1-\delta}{\delta}\right) + \delta \ln\left(\frac{\delta}{1-\delta}\right).$$

$$= (1-2\delta)\ln\left(\frac{1-\delta}{\delta}\right) \geq \ln\left(\frac{1}{4\delta}\right)$$

Let $\alpha_k = \dfrac{\mathbb{E}_\nu[N_a(\tau)]}{\mathbb{E}_\nu[\tau]}$ $\alpha \in \mathcal{P}_K$ and we have shown:

$$\mathbb{E}_\nu[\tau] \sum_{k=1}^{K} \alpha_k KL(\nu_k, \nu'_k) \geq \ln\left(\frac{1}{4\delta}\right).$$

for any $\nu' \in \mathcal{D}_{alt}(\nu)$ and a specific $\alpha \in \mathcal{P}_K$ that is independent of ν'

so: $$\mathbb{E}_\nu[\tau] \underbrace{\sup_{\alpha \in \mathcal{P}_K} \inf_{\nu' \in \mathcal{D}_{alt}(\nu)} \sum_{k=1}^{K} \alpha_k KL(\nu_k, \nu'_k)}_{c^*(\nu)^{-1}} \geq \ln\left(\frac{1}{4\delta}\right).$$

$\square$

Stop and track algorithm

Idea of the algorithm is to track the lower bound and stop when τ is larger than the estimated lower bound.

$\alpha^*_k(\nu)$ corresponds to the proportion of pulls on k.

i.e. we should stop when

$$\underbrace{\inf_{\nu' \in \mathcal{D}_{alt}(\nu)} \sum_{k=1}^{K} N_k(t)\, KL(\nu_k, \nu'_k)}_{Z_t} \geq \ln\left(\frac{1}{\delta}\right).$$

Problem: ν is unknown, but can be estimated.

Assume in the following $\mathcal{D} = \{ \mathcal{N}(\mu, 1) \mid \mu \in \mathbb{R} \}$

In that case, we can approximate Z_t by

$$Z_t := \boxed{\frac{1}{2} \inf_{\mu' \in M_{alt}(\hat{\mu})} \sum_{k=1}^{K} N_k(t) \left(\hat{\mu}_k(t) - \mu' \right)^2 .}$$

$$M_{alt}(\hat{\mu}) = \left\{ \mu' \mid \arg\max_k \hat{\mu}_k \cap \arg\max_k \mu'_k = \emptyset \right\}$$

Track-and-stop algorithm Input δ and $\beta_t(\delta)$

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

Pull $a_t = t$.

While $Z_t < \beta_t(\delta)$

if $\min_k N_k(t) < \sqrt{t}$ then pull $a_{t+1} \in \arg\min_k N_k(t)$ forced exploration

else choose $a_{t+1} \in \arg\max_k t\hat{\alpha}_k(t) - N_k(t)$ track

stop and return $\psi \in \arg\max_k \hat{\mu}_k(t)$ stop.

$$\boxed{\hat{\alpha}(t) \in \arg\max_{\alpha \in P_K} \inf_{\nu' \in \mathcal{D}_{alt}(\hat{\nu}(t))} \sum_{k=1}^{K} \alpha_k \, KL(\hat{\nu}(t), \nu')}$$

For our Gaussian setting,

Theorem There exists a choice $\beta_t(\delta)$ such that track-and-stop is sound for the Gaussian setting and for any ν with a unique optimal arm:

$$\lim_{\delta \to 0} \frac{E[\tau]}{\ln(1/\delta)} = c^*(\nu).$$

Lemma: let $f : [K, +\infty) \to \mathbb{R}$ be given by

$f(x) = \exp(K - x) \left(\frac{x}{K}\right)^K$ and $\beta_t(\delta) = K \ln(t^2 + t) + f^{-1}(\delta)$.

Then for $\tau = \min \{ t \mid Z_t \geq \beta_t(\delta) \}$, it holds $P\left(\operatorname{argmax}_k \hat{\mu}_k(t) \neq \operatorname{argmax}_k \mu_k\right) \leq \delta$.

Sketch of proof of the Theorem

(detailed proof in exercise session #6)

Taking $\beta_t(\delta)$ as in the lemma, the strategy is sound.

The idea of the proof is to show that we have: $\hat{\mu}(t) \approx \mu$ and $\frac{N_k(t)}{t} \approx \hat{\alpha}(t) \approx \alpha^*(\nu)$

From there:

$$Z_t = \frac{1}{2} \inf_{\mu' \in \text{Alt}(\hat{\mu}(t))} \sum_{k=1}^{K} N_k(t) \left(\hat{\mu}_k(t) - \mu_k'\right)^2$$

$$\approx t \inf_{\mu' \in \text{Alt}(\mu)} \sum_{k=1}^{K} \frac{\alpha_k^*(\nu) (\mu_k - \mu_k')^2}{2}$$

$$= \frac{t}{c^*(\nu)}.$$

The algorithm should then stop when $\frac{t}{c^*(\nu)} \gtrsim \beta_t(\delta) = (1 + o(1)) \ln(1/\delta)$.

i.e. $\tau \sim \ln(1/\delta)$.

Proof of the lemma:

If $|\arg\max_k \hat{\mu}_k(t)| > 1$, then $Z_t = 0$. So assume in the following $\tau < \infty$ and

$\underbrace{|\arg\max_k \hat{\mu}_k(\tau)| = 1}$.

denote ψ

By definition of $\mathcal{M}_{alt}(\hat{\mu})$, for any $k^* \in \arg\max_k \mu_k$

$$\mathbb{P}(k^* \neq \psi \text{ and } \tau<\infty) = \mathbb{P}\left(\mu \in \mathcal{M}_{alt}(\hat{\mu}(\tau)) \text{ and } \tau < \infty\right)$$

By def^n of Z_t, for any $\mu' \in \mathcal{M}_{alt}(\hat{\mu}(\tau))$: $\frac{1}{2}\sum_{k=1}^{K} N_k(\tau)\left(\hat{\mu}_k(\tau) - \mu'_k\right)^2 \geq \beta_\tau(\delta)$, so

$$\mathbb{P}(k^* \neq \psi \text{ and } \tau < \infty) \leq \mathbb{P}\left(\frac{1}{2}\sum_{k=1}^{K} N_k(\tau)\left(\hat{\mu}_k(\tau) - \mu_k\right)^2 \geq \beta_\tau(\delta)\right).$$

$(\beta_t(\delta) = K\ln(t^2+t) + f^{-1}(\delta))$

We can show the two following concentration inequalities, which allow to conclude

(1) $\mathbb{P}\left(\exists t \geq 1, \frac{N_k(t)}{2}\left(\hat{\mu}_k(t) - \mu_k\right)^2 \geq \ln\left(\frac{1}{\delta}\right) + \ln(t^2+t)\right) \leq \delta$

(2) $\mathbb{P}\left(\exists t \geq 1, \sum_{k=1}^{K} \frac{N_k(t)}{2}\left(\hat{\mu}_k(t) - \mu_k\right)^2 \geq K\ln(t^2+t) + x\right) \leq \left(\frac{x}{K}\right)^K \exp(K - x)$

Proofs are omitted.
Easier to prove in the alternative model where $Y(t) = X_{a_t}(N_{a_t}(t))$.

$\square$

Setting 2: (fixed budget)

At each round $t = 1, \dots, T$:

- agent picks an arm $a_t \in [K]$ (based on previous observations)
- observes $X_{a_t}(t) \sim \nu_{a_t} \in \mathcal{D}$

After round T, return $\psi \in [K]$.

Goal: maximize $\mathbb{P}(\psi \in \arg\max_k \mu_k)$

→ much harder problem