Lecture #6 : Lower bounds and Best arm Identification

Lower bound

In lecture 3, we proposed algorithms with (pseudo) regrets bounded as

$$R_T \leqslant c \sum_{k, \Delta_k > 0} \frac{\ln T}{\Delta_k}$$ (instance dependent regret)

Is it possible to do better?

Our goal: show that for "good" algorithms π, $\mathbb{E}[R_T(\nu, \pi)] \geqslant f(\pi, \nu)$ for some lower bound f.

Before that, we need to introduce some information theory tools.

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

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

$\mathbb{Q}(A) = 0 \Rightarrow \mathbb{P}(A) = 0$

First properties:

- $KL(\mathbb{P}, \mathbb{Q}) \geqslant 0$ (by concavity of the log)
- Joint convexity: for $\lambda \in [0,1]$ and $\mathbb{P}_1, \mathbb{P}_2, \mathbb{Q}_1, \mathbb{Q}_2$:

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

- Not a distance: not symmetric and no triangle inequality

Theorem (data processing inequality)

Let $X, Y \in \mathcal{X}$ be random variables, let $U \in \mathcal{U}$ be a r.v. independent from X, Y and let $\varphi : \mathcal{X} \times \mathcal{U} \to \mathcal{Z}$ be a measurable function. Then

$$KL(\varphi(X,U), \varphi(Y,U)) \leq KL(X,Y)$$

(we write $KL(X,Y)$ for the KL between the distributions of X and Y)

"Processing" random variables can only lose information and make them closer in KL.

Define $\mathbb{P}^X$ the law of X under $\mathbb{P}$

and $KL(\mathbb{P}^{X|Y}, \mathbb{Q}^{X|Y}) = \mathbb{E}_{y \sim \mathbb{P}^Y}\left[KL(\mathbb{P}^{X|Y=y}, \mathbb{Q}^{X|Y=y})\right]$

Chain rule for KL

$$KL(\mathbb{P}^{(X,Y)}, \mathbb{Q}^{(X,Y)}) = KL(\mathbb{P}^{X|Y}, \mathbb{Q}^{X|Y}) + KL(\mathbb{P}^Y, \mathbb{Q}^Y)$$

Let $H_t = (U_0, X_{a_1}(1), U_1, \dots, X_{a_t}(t), U_t)$ be the history until time t (r.v.s U account for potential randomisation of algorithm)

A policy π is a sequence of measurable functions π_t s.t.

$$\pi_{t+1}(H_t) = a_{t+1}$$

Lemma: (fundamental inequality for stochastic bandits)

For two different bandit instances $\nu = (\nu_k)_{k \in [K]}$ and $\nu' = (\nu'_k)_{k \in [K]}$,

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

$$\sum_{k=1}^{K} \mathbb{E}_\nu[N_k(T)]\, KL(\nu_k, \nu'_k) = KL(\mathbb{P}_\nu^{H_T}, \mathbb{P}_{\nu'}^{H_T})$$
$$\geq KL(\mathrm{Ber}(\mathbb{E}_\nu[Z]), \mathrm{Ber}(\mathbb{E}_{\nu'}[Z]))$$

Proof: • the $\geq$ is a direct consequence of the processing inequality

with $\varphi(H_T, U) = \mathbb{1}(U \leq Z)$.

• The equality comes from the chain rule:

$$KL(\mathbb{P}_\nu^{H_{t+1}}, \mathbb{P}_{\nu'}^{H_{t+1}}) = KL(\mathbb{P}_\nu^{U_{t+1}, X_{a_{t+1}}(t+1) | H_t}, \mathbb{P}_{\nu'}^{U_{t+1}, X_{a_{t+1}}(t+1) | H_t}) + KL(\mathbb{P}_\nu^{H_t}, \mathbb{P}_{\nu'}^{H_t})$$

$$= \mathbb{E}\Big[\sum_{k=1}^K \mathbb{1}\{k = a_{t+1}\} KL(\mathbb{P}_\nu^{X_k(t+1)}, \mathbb{P}_{\nu'}^{X_k(t+1)})\Big] + KL(\mathbb{P}_\nu^{H_t}, \mathbb{P}_{\nu'}^{H_t})$$

$$= \sum_{k=1}^K \mathbb{E}[\mathbb{1}\{k = a_{t+1}\}] KL(\nu_k, \nu'_k) + KL(\mathbb{P}_\nu^{H_t}, \mathbb{P}_{\nu'}^{H_t})$$

It then follows by induction. ☒

This lemma is key to proving the following lower bound.

Defn

A strategy is consistent w.r.t a model $\mathcal{D}$ if, for all bandit instances $\nu \in \mathcal{D}^K$, $\forall \alpha \in (0,1]$, $\forall k$ s.t. $\Delta_k > 0$,

$\mathbb{E}[N_k(T)] = o(T^\alpha)$.

- Typical bounds for good strategies: $\forall \nu \in \mathcal{D}^K$, $\forall k$ s.t. $\Delta_k > 0$, $\limsup_{T \to \infty} \frac{\mathbb{E}[N_k(T)]}{\ln T} \leq C_k(\nu)$

(remember UCB)

- optimal such term is $C_k(\nu) = \frac{1}{K_{inf}(\nu_k, \mu^*; \mathcal{D})}$

where $K_{inf}(\nu_k, \mu^*; \mathcal{D}) = \inf \left\{ KL(\nu_k, \nu'_k) \;\middle|\; \begin{array}{l} \nu'_k \in \mathcal{D} \\ \mathbb{E}(\nu'_k) > \mu^* \end{array} \right\}$

We will only prove the lower bound for this term.

Theorem (Lai and Robbins, 1985, Burnetas and Katehakis, 1996)

For all bandit models $\mathcal{D} \in \mathcal{P}_1(\mathbb{R})$,
for any consistent strategy w.r.t. $\mathcal{D}$,
for any bandit instance $\nu \in \mathcal{D}^K$,
for all suboptimal arms k (ie $\Delta_k > 0$), $\liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\ln T} \geq \frac{1}{K_{inf}(\nu_a, \mu^*, \mathcal{D})}$

Corollary

for all bandit models $\mathcal{D}$, any consistent strategy wrt $\mathcal{D}$, all bandit instances $\nu \in \mathcal{D}^K$:

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

Proof:

$$K_{inf}(\nu_k, \mathcal{D}, \mu^*) = \inf\left\{ KL(\nu_k, \nu_k') \mid \nu_k' \in \mathcal{D},\ \nu_k \ll \nu_k' \text{ and } \mathbb{E}(\nu') > \mu^* \right\}.$$

Convention $\inf \emptyset = +\infty$

For the proof, we
- fix $\mathcal{D}$, strategy π, ν and k s.t. $\Delta_k > 0$
- fix an alternative model ν' with

$$\begin{cases} \nu_i' = \nu_i & \text{for all } i \neq k \\ \nu_k' \text{ s.t. } & \nu_k' \in \mathcal{D},\ \nu_k \ll \nu_k' \text{ and } \mathbb{E}(\nu_k') > \mu^* \end{cases}$$

That is ν and ν' only differ at k, the unique optimal arm in ν'.

We are using the fundamental inequality with

$$Z = \frac{N_k(T)}{T} \quad \text{which is } [0,1]\text{-valued}$$
$$\sigma(H_T)\text{-measurable}$$

The fundamental inequality (lemma) yields, since ν and ν' only differ at k:

$$\mathbb{E}_\nu[N_k(T)]\, KL(\nu_k, \nu_k') \geq KL\left(Ber\left(\mathbb{E}_\nu\left[\frac{N_k(T)}{T}\right]\right), Ber\left(\mathbb{E}_{\nu'}\left[\frac{N_k(T)}{T}\right]\right)\right)$$

$$\geq -\ln(2) + \left(1 - \mathbb{E}_\nu\left[\frac{N_k(T)}{T}\right]\right) \ln\left(\frac{1}{1 - \mathbb{E}_{\nu'}\left[\frac{N_k(T)}{T}\right]}\right)$$

indeed $KL(Ber(p), Ber(q)) = p \ln\left(\frac{p}{q}\right) + (1-p)\ln\left(\frac{1-p}{1-q}\right)$

$$= p \underbrace{\ln\left(\frac{1}{q}\right)}_{\geq 0} + (1-p)\ln\left(\frac{1}{1-q}\right) + \underbrace{\left(p\ln(p) + (1-p)\ln(1-p)\right)}_{\geq -\ln 2}$$

$$\geq -\ln 2 + (1-p)\ln\left(\frac{1}{1-q}\right) \quad \text{for all } (p,q) \in [0,1) \quad (\text{and even for } p,q \in [0,1])$$

π is consistent, so:

- instance ν → k is suboptimal $\mathbb{E}_\nu\left[\frac{N_k(T)}{T}\right] \xrightarrow[T\to\infty]{} 0$

- instance ν' → all $i \neq k$ are suboptimal:

  for any $\alpha \in]0,1]$, $\mathbb{E}_{\nu'}[N_i(T)] = o(T^\alpha)$

In particular: $T - \mathbb{E}_{\nu'}[N_k(T)] = \sum_{i\neq k} \mathbb{E}_{\nu'}[N_i(T)] = o(T^\alpha)$

so:
$$\frac{1}{1-\mathbb{E}_{\nu'}\left[\frac{N_a(T)}{T}\right]} = \frac{T}{T-\mathbb{E}_{\nu'}[N_a(T)]} = \frac{T}{o(T^{\alpha})}$$

$$\geqslant T^{1-\alpha} \quad \text{for } T \text{ large enough}$$

substituting back and dividing by $\ln T$: for any $\alpha \in (0,1)$ and T large enough

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

thus $$\liminf_{T\to+\infty} \frac{\mathbb{E}_\nu[N_a(T)]}{\ln T} \geqslant \frac{(1-\alpha)}{KL(\nu_a,\nu_a')}$$ (true whether the KL is $<+\infty$ or $=+\infty$ (it is necessarily >0))

for any $\alpha \in (0,1)$, so $$\liminf_{T\to+\infty} \frac{\mathbb{E}_\nu[N_a(T)]}{\ln T} \geqslant \frac{1}{KL(\nu_a,\nu_a')}$$

Holds for any $\nu_a' \in \mathcal{D}$ s.t. $\nu_a \ll \nu_a'$ and $E(\nu_a') > \mu^*$, so that taking the supremum of the right hand side on these ν_a' yields the lower bound:

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

Comments on the lower bound:

- algorithms with optimal instance dependent bounds are know (e.g. KL-UCB, Thompson sampling) but require a long and technical analysis.

- this is an asymptotic lower bound for $T \to +\infty$.
  what about small T?

Theorem (minimax lower bound)

Let $\mathcal{D} = \{ \mathcal{N}(\mu, 1) \mid \mu \in \mathbb{R} \}$, $K \geq 2$ and $T \geq K-1$

There exists a universal constant $c > 0$ such that,

for any policy π, there ~~exists~~ $\nu \in \mathcal{D}^K$ s.t.

$$R_T(\pi, \nu) \geq c \sqrt{KT}$$

- Case 1: $\mathcal{D} = \{ \mathcal{N}(\mu, \sigma^2) \mid \mu \in \mathbb{R} \}$

then

$$K_{\inf}(\nu_k, \mu^*, \mathcal{D}) = \frac{\Delta_k^2}{2\sigma^2}$$

Best possible regret of order $2\sigma^2 \sum_{k, \Delta_k > 0} \frac{\ln T}{\Delta_k}$

UCB has regret $\leq 32\sigma^2 \sum_{k, \Delta_k > 0} \frac{\ln T}{\Delta_k}$

↳ optimal up to constant factor
can be made optimal with finer version

- <u>Case 2</u> : $\mathcal{D} = \{ Ber(p) \mid p \in [0,1] \}$

then

$$K_{inf}(\nu_k, \mu^*, \mathcal{D}) = \mu_k \ln \frac{\mu_k}{\mu^*} + (1-\mu_k) \ln \frac{1-\mu_k}{1-\mu^*}$$

Best arm identification

<u>Until now:</u> 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.

Example: simple regret minimisation: $\Delta_{T+1} = \mu^* - \mu_{a_{T+1}}$

Setting 1: BAI, 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

Question: which arm has the highest mean?

confidence level

Goal: 1) Have a δ-correct strategy: $\mathbb{P}(\tau < \infty \text{ and } \mu_\psi < \mu^*) \leq \delta$ (for any ν)

with confidence level $\delta \in (0,1)$

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

Theorem (lower bound)

Let (π, τ, ψ) be a δ-correct strategy for the bandit model $\mathcal{D} = \{\mathcal{N}(\mu, 1) \mid \mu \in \mathbb{R}\}$ and let $\nu \in \mathcal{D}^K$. Then:

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

$$c^*(\nu)^{-1} = \sup_{\alpha \in \mathcal{P}_K} \left(\inf_{\mu' \in \mathrm{Alt}(\mu)} \sum_{k=1}^K \alpha_k \frac{(\mu_k - \mu'_k)^2}{2} \right)$$

where $\mathrm{Alt}(\mu) = \left\{ \mu' \in \mathbb{R}^K \mid \arg\max_k \mu_k \cap \arg\max_k \mu'_k = \emptyset \right\}$

i.e. no arm is optimal for both ν and ν'

Proof is similar to regret lower bound, but with ν' given by $\mu' \in Alt(\mu)$.

$$Z = \mathbb{1}_{\{\tau < \infty \text{ and } \psi \notin \arg\max_k \mathbb{E}(\nu'_k)\}}$$

Remarks

- α_k represents the "optimal" fraction of pulls on arm k.

  It indeed appears in the proof that $c^*(\nu) \geq \sum_{k=1}^{K} \mathbb{E}\left[\frac{N_k(\tau)}{\tau}\right] KL(\nu_k, \nu'_k)$

- This suggest an optimal non-adaptive sampling allocation:

$$N_k(\tau) = \tau \, \alpha^*_k \qquad \text{with } \alpha^* = \arg\max_{\alpha \in P_K} \left(\inf_{\mu' \in Alt(\mu)} \sum_{k=1}^{K} \alpha_k \frac{(\mu'_k - \mu_k)^2}{2} \right)$$

i.e. we should stop when

$$\underbrace{\inf_{\mu' \in Alt(\mu)} \sum_{k=1}^{K} N_k(t) \frac{(\mu_k - \mu'_k)^2}{2}}_{Z^*_t} \geq \ln\left(\frac{1}{\delta}\right)$$

Problem: μ is unknown, but can be estimated by $\hat{\mu}$

In that case, we can approximate Z^*_t by

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

Track-and-stop algorithm

Input δ and $\beta_t(\delta)$

Pull all arms once

While $Z_t < \beta_t(\delta)$

- if $\min_a N_a(t) \leq \sqrt{t}$ then pull $a_{t+1} \in \arg\min_a N_a(t)$ — forced exploration
- else choose $a_{t+1} \in \arg\max_a t\hat{\alpha}_a(t) - N_a(t)$ — track

Stop and return $\psi \in \arg\max_a \hat{\mu}_a(t)$ — stop.

Theorem

If $\nu \in D = \{\mathcal{N}(\mu, 1) \mid \mu \in \mathbb{R}\}$, Track-and-Stop is δ-correct and asymptotically optimal for $\beta_t(\delta) \approx K\ln(t) + \ln\left(\frac{1}{\delta}\right)$:

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_\mu[\tau]}{\ln\left(\frac{1}{\delta}\right)} \leq \frac{1}{\sup_{\alpha \in \mathcal{P}_K} \inf_{\mu' \in \mathrm{Alt}(\mu)} \sum_a \alpha_a \frac{(\mu_a - \mu'_a)^2}{2}}$$

Omitted proof relies on key concentration bounds

Remarks

- Track-and-Stop is computationally intensive due to the oracle allocation computation ($\hat{\alpha}$)
- Forced exploration is wasteful in practice

  (we could use optimism instead)

- still no good bound for fixed $\delta > 0$ (e.g. $\delta = 0{,}05$)

Setting 2 : BAI, fixed budget

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

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

At time T: return a final choice $\psi \in [K]$

Goal: minimize $\mathbb{P}(\mu^* > \mu_\psi)$

Complexities: $H_1 = \sum_{k, \Delta_k > 0} \frac{1}{\Delta_k^2}$, $H_2 = \max_{k, \Delta_k > 0} \frac{k}{\Delta_{(k)}^2}$ reordered values: $\Delta_{(1)} \leq \Delta_{(2)} \leq \dots \leq \Delta_{(K)}$

$$H_2 \leq H_1 \leq \ln(2K) H_2$$

Lower bound of order
$$\begin{cases} \exp\left(-\frac{T}{\ln(K) H_1}\right) & \text{if } H_1 \text{ unknown} \\ \exp\left(-\frac{T}{H_1}\right) & \text{if known} \end{cases}$$

First approach: uniform exploration of the arms. Good baseline, but not very good when arms have very different means.

Sequential Halving:

Set $L = \lceil \log_2(K) \rceil$ and $A_1 = [K]$

For $\ell = 1, \dots, L$:

- Pull each arm in A_ℓ $T_\ell = \lfloor \frac{T}{L |A_\ell|} \rfloor$ times
- Let $\hat{\mu}_i^\ell$ be the empirical mean of arm i based only on these last T_ℓ samples
- Let $A_{\ell+1}$ contain the top $\lceil \frac{|A_\ell|}{2} \rceil$ arms in A_ℓ

Return ψ as the only arm in A_{L+1}

Thm:

If the distributions are 1-sub-Gaussian, then Sequential Halving satisfies:

$$\mathbb{P}(\mu^* > \mu_\psi) \leq 3 \log_2(K) \exp\left(- \frac{T}{16 H_2 \log_2(K)}\right).$$

Remarks

- close to lower bound
- For uniform exploration, we can bound this probability by

$$\sum_{k, \Delta_k > 0} \exp\left(- \frac{\lfloor T/K \rfloor \Delta_k^2}{4}\right).$$

- VE slightly better than SH when $\Delta_k = \Delta$ for any k $\quad (H_2 = \frac{K}{\Delta^2})$.

but SH much better than VE when $\Delta_2 = \Delta \ll 1$ $\quad (H_2 = \frac{1}{\Delta^2})$.

$\Delta_k = 1$ for $k > 2$