DATA 37200: Learning, Decisions, and Limits

(Winter 2026)

Lecture 4: Information-theoretic lower bounds

Instructor: Frederic Koehler

Reference

Up to some minor details, we will follow the proof in Tor Lattimore and Csaba Szepesvari's Bandits book, Chapters 13-15.

Goal: minimax lower bound for K-armed stochastic bandits

Setting: K arms, T rounds. At time t algorithm chooses arm $i(t) \in [K]$ and observes reward $r(t) \in [0, 1]$ with

$$\mathbb{E}[r(t) \mid i(t) = i] = \mu_i$$

Define pseudo-regret

$$\tilde{R}_T := \sum_{t=1}^{T} \left(\mu^\star - \mu_{i(t)}\right), \qquad \mu^\star := \max_i \mu_i,$$

(and similarly regret with realized rewards).

Theorem (minimax rate). There exists a universal constant $c > 0$ such that for any algorithm/policy π,

$$\sup_{\text{instances on } [0,1]} \mathbb{E}_\pi[\tilde{R}_T] \;\geq\; c\sqrt{KT}.$$

(whereas UCB1 showed an upper bound of $O(\sqrt{KT \log(T)})$).

Sidenote: minimax upper bound

The most well-known algorithm achieving the "minimax" upper bound $O(\sqrt{KT})$ is called "MOSS". Minimax here refers to the fact that:

$$\min_{\text{policies}} \max_{\text{instances on } [0,1]} \mathbb{E}[\tilde{R}_T] = \Theta(\sqrt{KT})$$

It shaves the $\sqrt{\log(T)}$ factor from UCB1.

9.1 The MOSS Algorithm

Algorithm 7 shows the pseudocode of MOSS, which is again an instance of the UCB family. The main novelty is that the confidence level is chosen based on the number of plays of the individual arms, as well as n and k.

Input n and k
Choose each arm once
Subsequently choose

$$A_t = \operatorname{argmax}_i \hat{\mu}_i(t-1) + \sqrt{\frac{4}{T_i(t-1)} \log^+\left(\frac{n}{kT_i(t-1)}\right)},$$

$$\text{where } \log^+(x) = \log \max\{1, x\} .$$

Algorithm 7: MOSS.

Key mathematical concept: KL Divergence

KL divergence is an important measure of *distance* between probability distributions which is used all the time in ML, RL, and many other areas. It is a concept core to information theory and statistics.

Although it is not required to deeply understand the KL divergence to go through the proof, it helps when a lot to understand its intuitive relation to Shannon's source coding theorem (i.e. mathematics of *optimal compression* for i.i.d. samples).

Motivation: Shannon entropy and optimal compression

Let X be a discrete random variable on alphabet $\mathcal{X}$ with distribution P.

Shannon entropy (bits):

$$H_2(P) := \sum_{x \in \mathcal{X}} P(x) \log_2 \frac{1}{P(x)}.$$

Source coding theorem (informal statement). For i.i.d. data $X_1, \ldots, X_n \sim P$:

- There exist lossless codes whose expected *per-symbol* length approaches $H_2(P)$ as $n \to \infty$. (see "Huffman code")
- No lossless code can achieve expected per-symbol length below $H_2(P)$ in the limit.

So $H_2(P)$ is the fundamental limit for compression: the best achievable average bits per symbol.

Cross-entropy and the definition of KL divergence

Suppose the data are truly $X \sim P$, but we encode as if $X \sim Q$.

The **cross-entropy** (bits) is

$$H_2(P, Q) := \sum_x P(x) \log_2 \frac{1}{Q(x)}.$$

Heuristically, this is the average code length if we use a code optimized for Q.

Compare to the optimal limit $H_2(P)$. Their difference is

$$H_2(P, Q) - H_2(P) = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)} =: D_{\mathrm{KL},2}(P \| Q).$$

Using natural logs gives the standard (nats) definition:

$$D_{\mathrm{KL}}(P \| Q) := \sum_x P(x) \log \frac{P(x)}{Q(x)}.$$

Interpretation: KL is the *redundancy* (extra bits/nats per symbol) from coding P-data with a Q-model.

Nonnegativity of KL

Assume $P \ll Q$ and write $p(x), q(x)$ for pmfs. Define the likelihood ratio

$$L(x) := \frac{p(x)}{q(x)}.$$

Then (a useful identity)

$$D_{\mathrm{KL}}(P\|Q) = \sum_x q(x)\,\frac{p(x)}{q(x)}\log\frac{p(x)}{q(x)} = \mathbb{E}_{X\sim Q}\big[\,f(L(X))\,\big],$$

where $f(u) = u\log u$.

Key facts: f is convex on $(0,\infty)$ since $f''(u) = 1/u > 0$, and

$$\mathbb{E}_Q[L(X)] = \sum_x q(x)\frac{p(x)}{q(x)} = \sum_x p(x) = 1.$$

By Jensen's inequality (convex f),

$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_Q[f(L)] \;\geq\; f(\mathbb{E}_Q[L]) = f(1) = 0.$$

Equality holds iff $L(X)$ is a.s. constant under Q, i.e. $P = Q$.

Chain rule for KL: statement

Let (X, Y) be discrete random variables. Write P_{XY} and Q_{XY} for two joint distributions on $\mathcal{X} \times \mathcal{Y}$, with marginals P_X, Q_X and conditionals $P_{Y|X}, Q_{Y|X}$.

Theorem (chain rule). Assuming $P_{XY} \ll Q_{XY}$,

$$D_{\mathrm{KL}}(P_{XY} \| Q_{XY}) = D_{\mathrm{KL}}(P_X \| Q_X) + \mathbb{E}_{X \sim P_X} D_{\mathrm{KL}}\big(P_{Y|X}(\cdot|X) \,\|\, Q_{Y|X}(\cdot|X)\big)$$

Expanded form:

$$D_{\mathrm{KL}}(P_{XY} \| Q_{XY}) = \sum_x P(x) \log \frac{P(x)}{Q(x)} + \sum_x P(x) \sum_y P(y|x) \log \frac{P(y|x)}{Q(y|x)}.$$

Chain rule for KL: proof (discrete)

Start from the definition and factor the joint laws:

$$D_{\mathrm{KL}}(P_{XY}\|Q_{XY}) = \sum_{x,y} P(x,y)\log\frac{P(x,y)}{Q(x,y)} = \sum_{x,y} P(x,y)\log\frac{P(x)P(y|x)}{Q(x)Q(y|x)}$$

Split the logarithm:

$$= \sum_{x,y} P(x,y)\log\frac{P(x)}{Q(x)} + \sum_{x,y} P(x,y)\log\frac{P(y|x)}{Q(y|x)}.$$

First term:

$$\sum_{x,y} P(x,y)\log\frac{P(x)}{Q(x)} = \sum_{x} P(x)\log\frac{P(x)}{Q(x)} = D_{\mathrm{KL}}(P_X\|Q_X).$$

Second term:

$$\sum_{x,y} P(x,y)\log\frac{P(y|x)}{Q(y|x)} = \mathbb{E}_{X\sim P_X}\left[D_{\mathrm{KL}}(P_{Y|X}\|Q_{Y|X})\right].$$

Chain rule: coding intuition

Coding a pair (X, Y) with model Q_{XY} has redundancy $D_{\mathrm{KL}}(P_{XY} \| Q_{XY})$.

Coding in two stages (first X using Q_X, then $Y|X$ using $Q_{Y|X}$) has redundancy

$$D_{\mathrm{KL}}(P_X \| Q_X) + \mathbb{E}_{X \sim P_X} \left[D_{\mathrm{KL}}(P_{Y|X} \| Q_{Y|X}) \right],$$

which must equal the joint redundancy — this is exactly the chain rule.

KL and a testing inequality

Now we understand

$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_P\left[\log \frac{dP}{dQ}\right]$$

so we can start to apply it.

Bretagnolle–Huber (proof left to HW). For any event A and any P, Q,

$$P(A) + Q(A^c) \;\geq\; \frac{1}{2}\exp\big(-D_{\mathrm{KL}}(P\|Q)\big).$$

Interpretation: if $D_{\mathrm{KL}}(P\|Q)$ is small, no test can separate P vs Q with small error.

We will apply this with $P =$ "rollout under instance 0" and $Q =$ "rollout under instance j".

Key lemma: chain rule for KL under adaptive sampling

Let $\nu = (D_1, \ldots, D_K)$ and $\nu' = (D'_1, \ldots, D'_K)$ be two bandit instances. Let P_ν and $P_{\nu'}$ be the induced distributions on the full transcript $\mathcal{F}_T = (i(1), r(1), \ldots, i(T), r(T))$ under a fixed algorithm.

Lemma (bandit KL decomposition).

$$D_{\mathrm{KL}}(P_\nu \| P_{\nu'}) \;=\; \sum_{i=1}^{K} \mathbb{E}_\nu[N_i(T+1)]\; D_{\mathrm{KL}}(D_i \| D'_i),$$

where $N_i(t) = \#\{s \le t-1 :\; i(s) = i\}$.

Proof: direct application of the chain rule for KL.

Hard instances: one slightly-better arm

Fix $\varepsilon \in (0, 1/4]$.

Define instance ν^0: all arms are $\mathrm{Ber}(1/2)$.

For each $j \in [K]$, define instance ν^j:

$$D_j = \mathrm{Ber}(1/2 + \varepsilon), \qquad D_i = \mathrm{Ber}(1/2) \;\; (i \neq j).$$

Then $\mu^\star = 1/2 + \varepsilon$ and the unique best arm is j.

Regret on ν^j. Each time we do *not* play j, we lose gap ε, so

$$\mathbb{E}_{\nu^j}[\tilde{R}_T] = \varepsilon\, \mathbb{E}_{\nu^j}\big[T - N_j(T+1)\big].$$

So it suffices to show that for some j the algorithm fails to pull arm j often.

Warmup: $\Omega(\sqrt{T})$ bound

Because

$$D_{KL}(Ber(1/2)\|Ber(1/2+\epsilon)) = D_{KL}(Ber(1/2)\|Ber(1/2-\epsilon))$$

and because it is analytic, we know that

$$D_{KL}(Ber(1/2)\|Ber(1/2+\epsilon)) \lesssim \epsilon^2$$

for small ϵ. (Why?)
Consider a bandit with two arms labeled 1 and 2, and consider reward distributions ν^1 and ν^2 from before. By the bandit KL decomposition and the fact that no arm is pulled more than T times, we have

$$D_{KL}(P_{\nu^1}, P_{\nu^2}) \lesssim T\epsilon^2.$$

If $\epsilon \ll 1/\sqrt{T}$, then by Bretagnolle-Huber we either pull arm 1 a lot under ν^2 or pull arm 2 a lot under ν^2, so we experience $\Omega(\epsilon)$ regret.

Next: $\Omega(\sqrt{KT})$ bound

Our lower bound in the previous slide clearly does not improve as the number of arms K increases. However, our best upper bound is $O(\sqrt{KT})$. It turns out we can improve the lower bound with a more sophisticated argument.

Pick an arm j that is rarely sampled under ν^0

Under ν^0, the algorithm still makes exactly T pulls total:

$$\sum_{i=1}^{K} N_i(T+1) = T \quad \Rightarrow \quad \sum_{i=1}^{K} \mathbb{E}_{\nu^0}[N_i(T+1)] = T.$$

Therefore, there exists some $j \in [K]$ with

$$\mathbb{E}_{\nu^0}[N_j(T+1)] \leq \frac{T}{K}.$$

Define the event

$$A := \left\{ N_j(T+1) > \frac{4T}{K} \right\}.$$

By Markov's inequality,

$$\Pr_{\nu^0}(A) \;\leq\; \frac{\mathbb{E}_{\nu^0}[N_j(T+1)]}{4T/K} \;\leq\; \frac{1}{4}.$$

So, under ν^0, with probability at least $3/4$ the algorithm does *not* pull arm j more than $4T/K$ times.

Bound distinguishability: $D_{\mathrm{KL}}(P_{\nu^0} \| P_{\nu^j})$

Let P_0 and P_j denote the transcript laws under ν^0 and ν^j.

By the KL decomposition lemma (previous slide) and since only arm j differs,

$$D_{\mathrm{KL}}(P_0 \| P_j) = \mathbb{E}_{\nu^0}[N_j(T+1)] \cdot D_{\mathrm{KL}}\big(\mathrm{Ber}(1/2) \,\|\, \mathrm{Ber}(1/2+\varepsilon)\big).$$

For $\varepsilon \le 1/4$, a crude bound for Bernoulli KL is

$$D_{\mathrm{KL}}\big(\mathrm{Ber}(1/2) \,\|\, \mathrm{Ber}(1/2+\varepsilon)\big) \;\le\; 8\varepsilon^2.$$

Therefore, using $\mathbb{E}_{\nu^0}[N_j(T+1)] \le T/K$,

$$D_{\mathrm{KL}}(P_0 \| P_j) \;\le\; 8\varepsilon^2 \, \frac{T}{K}.$$

Apply Bretagnolle–Huber to go to ν^j

Recall $A = \{N_j(T+1) > 4T/K\}$ and $\Pr_{\nu^0}(A) \le 1/4$.

Apply Bretagnolle–Huber with $P = P_0$, $Q = P_j$:

$$\Pr_{\nu^0}(A) + \Pr_{\nu^j}(A^c) \;\ge\; \frac{1}{2}\exp\big(-D_{\mathrm{KL}}(P_0 \| P_j)\big).$$

Hence

$$\Pr_{\nu^j}(A^c) \;\ge\; \frac{1}{2}\exp\Big(-8\varepsilon^2 \frac{T}{K}\Big) \;-\; \frac{1}{4}.$$

Choose $\varepsilon := \sqrt{\frac{K}{32T}}$ so that $8\varepsilon^2 \frac{T}{K} = \frac{1}{4}$. Then

$$\Pr_{\nu^j}\left(N_j(T+1) \le \frac{4T}{K}\right) = \Pr_{\nu^j}(A^c) \;\ge\; \frac{1}{2}e^{-1/4} - \frac{1}{4} \;\ge\; c_0$$

for a universal constant $c_0 > 0$.

Conclude $\Omega(\sqrt{KT})$

On the event $A^c = \{N_j(T+1) \le 4T/K\}$,

$$T - N_j(T+1) \;\ge\; T\left(1 - \frac{4}{K}\right) \;\ge\; \frac{T}{2} \qquad (\text{for } K \ge 8).$$

Therefore,

$$\mathbb{E}_{\nu^j}[\tilde{R}_T] = \varepsilon\, \mathbb{E}_{\nu^j}[T - N_j(T+1)] \;\ge\; \varepsilon \cdot \frac{T}{2} \cdot \Pr_{\nu^j}(A^c) \;\ge\; c_1\, \varepsilon T.$$

With $\varepsilon = \sqrt{\frac{K}{32T}}$, this gives

$$\mathbb{E}_{\nu^j}[\tilde{R}_T] \;\ge\; c\sqrt{KT}.$$

Remarks: (i) constants and the $K \ge 8$ condition can be handled cleanly by casework; (ii) since $|R_T - \tilde{R}_T| = O(\sqrt{T})$ by Azuma, the same minimax rate holds for realized regret.