Sequential learning – Lesson 6
Lower Bounds / Best Arm Identification

Rémy Degenne

February 17, 2023

INRIA Lille-Nord Europe

Stochastic bandit

At each time step $t = 1, \ldots, T$

- the player chooses an arm $k_t \in \Theta$ (compact decision/parameter set, often $\{1, \ldots, K\}$);
- the player observes the reward of the chosen arm only: $X_t^{k_t} \sim \nu_{k_t}$ $\rightarrow$ bandit feedback.

The goal of the player is to maximize their cumulative reward.

The main reference:

Tor Lattimore and Csaba Szepesvári, Bandit algorithms. Cambridge University Press, 2020.

(online on Tor Lattimore's webpage)

Objective: Regret Minimization

Goal until now: minimize the "cumulated" (pseudo)-regret: sum over all rounds.

$$R_T = T\mu^* - \sum_{t=1}^{T} \mu_{k_t} \, .$$

Strategy: exploit to minimize the current regret (based on past information) or explore to gain more info.

This Lecture

- Lower bounds on the regret.
  Does the regret have to be $O(\log(T)/\Delta)$? Is UCB a good algorithm?

- Best arm identification: a pure exploration task.
  What if we don't want to minimize the regret, but want to know the identity of the arm with highest mean?

UCB

Finitely many arms, no contexts.

Initialization For rounds $t = 1, \dots, K$ pull arm $k_t = t$.

For $t = K+1, \dots, T$, choose

$$k_t \in \operatorname*{arg\,max}_{k \in [K]} \left\{ \widehat{\mu}_{t-1}^{k} + \sqrt{\frac{2 \log t}{N_{t-1}^{k}}} \right\},$$

and get reward $X_t^{k_t}$.

UCB Regret Bounds

Theorem 1

If the distributions ν_k have supports all included in $[0,1]$ then for all k such that $\Delta_k > 0$

$$\mathbb{E}\left[N_T^k\right] \leqslant \frac{8\log T}{\Delta_k^2} + 2\,.$$

In particular, this implies that the expected regret of UCB is upper-bounded as

$$\mathbb{E}[R_T] \leqslant 2K + \sum_{k:\Delta_k>0} \frac{8\log T}{\Delta_k}\,.$$

Remarks :

- we can also prove $\mathbb{E}[R_T] \lesssim \sqrt{KT\log(T)}$. Close to the optimal $O(\sqrt{KT})$.
- Deals with multiple gaps, without any knowledge of the gaps.
- Anytime algorithm: does not depend on T.

Lower Bounds

Best Arm Identification

Why can't we have zero regret?

Consider bandit problem with means $\mu = (\mu_1, \dots, \mu_K)$ with $\mu_1 = \max_k \mu_k$.

Algorithm: pull only arm 1. $\rightarrow$ zero regret!

But linear regret if 1 is not the best arm.

Theorem 2 (Regret of UCB)

If the distributions ν_k have supports all included in $[0,1]$ then for all k such that $\Delta_k > 0$

$$\mathbb{E}[R_T] \leqslant 2K + \sum_{k:\Delta_k>0} \frac{8\log T}{\Delta_k}.$$

$\rightarrow$ low regret for all distributions with support in $[0,1]$.

What makes the deterministic algorithm or FTL bad

If an algorithm does not pull arm k, it has no information on its mean μ_k.

$\Rightarrow$ it cannot exclude the possibility that $\mu_k = \max_j \mu_j$.

$\Rightarrow$ it may think that $* = 1$ in a problem in which $* = j$ and get high regret.

The algorithm needs to explore all arms to distinguish the current bandit problem from alternatives in which the best arm is different.

Our goal in this section: show that a "good" algorithm has to pull all arms. We want "good" $\Rightarrow$ $\mathbb{E}[N_T^k] \geq f(T, \Delta)$.

Result

Definition (Asymptotically correct)

An algorithm for stochastic bandit regret minimization is said to be asymptotically correct if for all $\nu = (\nu_1, \dots, \nu_k)$ with supports in $[0, 1]$,

$$\mathbb{E}_\nu[R_T] = o(T^\alpha) \text{ for all } \alpha > 0 \ .$$

Theorem 3

For all asymptotically correct algorithms, for all arms k with $\Delta_k > 0$,

$$\liminf_{T \to +\infty} \frac{\mathbb{E}_\nu[N_T^k]}{\log T} \geq \frac{1}{\inf\{KL(\nu_k, \nu') | \mathbb{E}_{X \sim \nu'}[X] > \mu^*\}} \ .$$

where $\mu^* = \max_k \mathbb{E}_{X \sim \nu_k}[X]$.

Regret lower bound: $\liminf_{T \to +\infty} \frac{\mathbb{E} R_T}{\log T} \geq \sum_{k : \Delta_k > 0} \frac{\Delta_k}{\inf\{KL(\nu_k, \nu') | \mathbb{E}_{X \sim \nu'}[X] > \mu^*\}}$.

Definition (Absolute continuity)

A probability measure ν is said to be absolutely continuous with respect to another probability measure ν', which we denote by $\nu \ll \nu'$, if for all events A, $\nu'(A) = 0 \;\Rightarrow\; \nu(A) = 0$.

Definition (Kullback-Leibler divergence)

The Kullback-Leibler divergence (or relative entropy) of distribution ν with respect to ν' is defined as

$$KL(\nu, \nu') = \begin{cases} \int_\Omega \log \frac{d\mathbb{P}_\nu}{d\mathbb{P}_{\nu'}}(\omega) d\mathbb{P}_\nu(\omega) & \text{if } \nu \ll \nu' \\ +\infty & \text{otherwise} \end{cases},$$

where $\frac{d\mathbb{P}_\nu}{d\mathbb{P}_{\nu'}}$ is the Radon–Nikodym derivative of ν with respect to ν'.

Kullback-Leibler Divergence

$$KL(\nu, \nu') = \begin{cases} \mathbb{E}_{X\sim\nu}\left[\log \frac{d\mathbb{P}_\nu}{d\mathbb{P}_{\nu'}}(X)\right] & \text{if } \nu \ll \nu' \\ +\infty & \text{otherwise} \end{cases},$$

Properties:

- KL is non-negative (proved using the concavity of the log).
- KL is jointly convex: For $\lambda \in [0, 1]$ and $\nu_1, \nu_2, \nu_1', \nu_2'$,
  $KL(\lambda\nu_1 + (1-\lambda)\nu_2, \lambda\nu_1' + (1-\lambda)\nu_2') \le \lambda KL(\nu_1, \nu_1') + (1-\lambda)KL(\nu_2, \nu_2')$.
- KL is not symmetric and does not verify the triangle inequality. It is not a distance.

The Kullback-Leibler divergence is our measure of how much ν appears different from ν' when sampling from ν.

Lower bound

We want to show asymptotically correct $\Rightarrow \mathbb{E}[N_T^k] \geq f(T, \Delta)$.

Asymptotically correct $\implies$ the algorithm pulls different arms on instances where the best arm is different.

Main idea: to have a different behavior on these instances, the algorithm needs to distinguish the current ν from any ν' with different best arm. Let $H_{t,\nu} = (k_1, X_1, \dots, k_t, X_t)$ be the random variable that stores the history until time t when the rewards have distributions $(\nu_1, \dots, \nu_K)$.

$\rightarrow KL(H_{T,\nu}, H_{T,\nu'})$ has to be large enough.

Can we compute that KL?

Data processing inequality

Theorem 4 (Data processing inequality)

Let $X, Y \in \mathcal{X}$ be random variables, let $U \in \mathcal{U}$ be independent of X and 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.

Let $H_{t,\nu} = (k_1, X_1, \ldots, k_t, X_t)$ be the random variable that stores the history until time t when the rewards have distributions $(\nu_1, \ldots, \nu_K)$. Let $Z_{k,\nu} = \mathbb{I}\{U \leq \frac{N_T^k}{T}\}$ be a Bernoulli random variable with value 1 if U with uniform distribution is smaller than the fraction of pulls $\frac{N_T^k}{T}$. Then

$$KL(H_{t,\nu}, H_{t,\nu'}) \geq KL(Z_{k,\nu}, Z_{t,\nu'}) = KL(\mathcal{B}(\mathbb{E}_\nu[\frac{N_T^k}{T}]), \mathcal{B}(\mathbb{E}_{\nu'}[\frac{N_T^k}{T}]))$$

Chain rule for KL

Define $KL((X|Y)_\nu, (X|Y)_{\nu'}) = \mathbb{E}_{y \sim \mathbb{P}^Y_\nu} \left[KL((X|Y=y)_\nu, (X|Y=y)_{\nu'}) \right]$.

Then we have the chain rule

$$KL((X,Y)_\nu, (X,Y)_{\nu'}) = KL((X|Y)_\nu, (X|Y)_{\nu'}) + KL(Y_\nu, Y_{\nu'}) \,.$$

Example for a Markov chain: $Z \to Y \to X$:

$$KL((X,Y,Z)_\nu, (X,Y,Z)_{\nu'}) = KL((X|Y)_\nu, (X|Y)_{\nu'}) + KL((Y|Z)_\nu, (Y|Z)_{\nu'}) + KL(Z_\nu, Z_{\nu'}) \,.$$

KL Divergence in a bandit problem

Two bandit problems with arm distributions given by ν and ν'.

$H_{t,\nu} = (k_1, X_1, \ldots, k_t, X_t)$: history until time t when the rewards have distributions given by $\nu = (\nu_1, \ldots, \nu_K)$.

Decision model $H_{t-1} \to k_t \to X_t$.

We get from the chain rule:

$$KL(H_{t,\nu}, H_{t,\nu'}) = \sum_k \mathbb{E}_\nu[N_t^k] KL(\nu_k, \nu'_k) \; .$$

Lower bound

We want to show asymptotically correct $\Rightarrow$ $\mathbb{E}[N_T^k] \geq f(T, \Delta)$.

Asymptotically correct implies that the algorithm pulls different arms on instances where $*$ is different.

Suppose that ν is such that $* = 1$ and ν' is such that there exists j with $\mu_j' > \mu_1$ and $\nu_k' = \nu_k$ for $k \neq j$.

We have shown

- $KL(H_{t,\nu}, H_{t,\nu'}) = \sum_k \mathbb{E}_\nu[N_t^k] KL(\nu_k, \nu_k')$,
- $KL(H_t^\nu, H_t^{\nu'}) \geq KL(\mathcal{B}(\mathbb{E}_\nu[\frac{N_T^k}{T}]), \mathcal{B}(\mathbb{E}_{\nu'}[\frac{N_T^k}{T}]))$,

Then for our specific ν, ν',

$$\mathbb{E}_\nu[N_t^j] KL(\nu_j, \nu_j') = KL(H_{t,\nu}, H_{t,\nu'}) \geq KL(\mathcal{B}(\mathbb{E}_\nu[\frac{N_T^j}{T}]), \mathcal{B}(\mathbb{E}_{\nu'}[\frac{N_T^j}{T}])) \, .$$

Lower bound

$KL(\mathcal{B}(a), \mathcal{B}(b)) = a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b} \geq a \log \frac{1}{b} + (1-a) \log \frac{1}{1-b} - \log 2.$

$$\mathbb{E}_\nu[N_t^j] KL(\nu_j, \nu_j') \geq (1 - \mathbb{E}_\nu[\frac{N_T^j}{T}]) \log \frac{T}{T - \mathbb{E}_{\nu'}[N_T^j]} - \log 2 \, .$$

Now use the asymptotically correct hypothethis to get $\mathbb{E}_\nu[\frac{N_T^j}{T}] \to 0$ and $T - \mathbb{E}_{\nu'}[N_T^j] = o(T^\alpha)$ for all $\alpha > 0$. We obtain

$$\liminf_{T \to +\infty} \frac{\mathbb{E}_\nu[N_t^j] KL(\nu_j, \nu_j')}{\log T} \geq 1 \, .$$

This is valid for all ν' that differ from ν only on arm j, with $\mu_j' > \mu_1$, hence we can take a supremum over the inequalities obtained for each such ν'.

Results

Theorem 5

For all asymptotically correct algorithms, for all arms k with $\Delta_k > 0$,

$$\liminf_{T\to+\infty} \frac{\mathbb{E}_\nu[N_T^k]}{\log T} \geq \frac{1}{\inf\{KL(\nu_k, \nu'_k) | \mathbb{E}_{X\sim\nu'_k}[X] \geq \mu^*\}} .$$

Regret lower bound: $\liminf_{T\to+\infty} \frac{\mathbb{E}_\nu[R_T]}{\log T} \geq \sum_{k:\Delta_k>0} \frac{\Delta_k}{\inf\{KL(\nu_k,\nu'_k)|\mathbb{E}_{X\sim\nu'_k}[X]\geq\mu^*\}}$.

UCB upper bound: $\mathbb{E}[R_T] \lesssim \sum_{k:\Delta_k>0} \frac{\log T}{\Delta_k}$.

For Gaussians $\mathcal{N}(\cdot, 1)$: $KL(\nu_a, \nu_b) = \frac{1}{2}(\mu_a - \mu_b)^2$.

Improvements

The asymptotically correct condition can be replaced by a finite time condition.

Example: sub-UCB, if the regret verifies $\mathbb{E}[R_T] \leq C_1 \sum_k \frac{\log T}{\Delta_k} + C_2 \sum_k \Delta_k$.

The complexity term reflects prior information about the allowed distributions $\mathcal{M}$:

$$\inf\{KL(\nu_k, \nu'_k) | \nu'_k \in \mathcal{M}^k \wedge \mathbb{E}_{X \sim \nu'_k}[X] \geq \mu^*\}$$

Example: we may know that all arms have Bernoulli distributions.

Lower Bounds

Best Arm Identification

Identification

At each time step $t = 1, \dots, \tau$

- the player chooses an arm $k_t \in \Theta$ (compact decision/parameter set, often $\{1, \dots, K\}$);
- the player observes the reward of the chosen arm only: $X_t^{k_t} \sim \nu_{k_t}$;
- the player either stops or continues.

When the player stops: it returns an answer $\widehat{i}$.

Example: an arm, answer to the question "which arm has highest mean?"

The goal of the player is to return the correct answer with high probability, as soon as possible.

Goals

Question: which arm has highest mean?

The goal of the player is to return the correct answer with high probability, as soon as possible.

Several variables in what makes an algorithm good:

- τ: (random) time at which the algorithm stops.
- $\delta = \mathbb{P}(\widehat{i} \neq *)$: probability of mistake.

Possible settings:

- Fixed budget: for $\tau = T$ known beforehand, minimize δ.
- Fixed confidence: for fixed δ, ensure that $\mathbb{P}(\text{error}) \leq \delta$ and minimize τ.
  - minimize $\mathbb{E}[\tau]$ or
  - minimize T such that with probability $1 - \delta$, the algorithm is correct and $\tau \leq T$.

Fixed confidence BAI

Question: which arm has highest mean?

Task: <u>sample arms</u>, then decide to <u>stop</u> (stopping time τ) and <u>recommend</u> $\widehat{i} \in [K]$.

Requirement: $\mathbb{P}(\tau < +\infty \wedge \widehat{i} = *) \geq 1 - \delta$.

Goal: minimize $\mathbb{E}[\tau]$, expected sample complexity.

Lower bound on bandits with Gaussian distributions. Distributions $\mathcal{N}(\mu_k, 1)$.

For $\mu \in \mathbb{R}^K$, let $alt(\mu) = \{\lambda \in \mathbb{R}^K | *_\mu \notin \arg\max_k \lambda_k\}$.

Theorem 6

An algorithm which is δ-correct on all problems with Gaussian arms with variance 1 verifies for all $\mu \in \mathbb{R}^k$

$$\mathbb{E}_\mu[\tau] \geq \frac{KL(\mathcal{B}(\delta), \mathcal{B}(1-\delta))}{\sup_{w \in \triangle_K} \inf_{\lambda \in alt(\mu)} \sum_k w_k \frac{1}{2}(\mu_k - \lambda_k)^2} .$$

Optimal non-adaptive algorithm

Lower bound:

$$\mathbb{E}_\mu[\tau] \geq \frac{KL(\mathcal{B}(\delta), \mathcal{B}(1-\delta))}{\sup_{w \in \triangle_K} \inf_{\lambda \in alt(\mu)} \sum_k w_k \frac{1}{2}(\mu_k - \lambda_k)^2} \, .$$

This suggests an optimal (for that lower bound) non-adaptive sampling allocation:

$$N_{T,opt} = T w_{opt} = T \arg\max_{w \in \triangle_K} \inf_{\lambda \in alt(\mu)} \sum_k w_k \frac{1}{2}(\mu_k - \lambda_k)^2 \, .$$

Algorithm: Track and Stop

Idea: estimate the oracle allocation and follow it.

It we have an estimate $\widehat{\mu}_t \approx \mu$, then by a continuity argument

$w_{opt}(\widehat{\mu}_t) = \arg\max_{w \in \triangle_K} \inf_{\lambda \in alt(\mu)} \sum_k w_k \frac{1}{2}(\widehat{\mu}_{t,k} - \lambda_k)^2 \approx w_{opt}(\mu)$.

$\rightarrow$ if we make sure that $\widehat{\mu}_t \approx \mu$, then we can sample

$k_t = \arg\min N_t^k - t w_{opt}^k(\widehat{\mu}_t)$ (tracking).

Stopping rule

When can we stop?

An answer: when we have enough information to state that $\mu \notin alt(\mu)$ with high enough confidence.

How do we quantify that?

Generalized log-likelihood ratio:

$$LRT(\mu, \lambda, H_t) = \log \frac{d\mathbb{P}_\mu}{d\mathbb{P}_\lambda}(H_t) = \sum_{s=1}^{t} \log \frac{d\mathbb{P}_{\mathcal{N}(\mu_{k_s},1)}}{d\mathbb{P}_{\mathcal{N}(\lambda_{k_s},1)}}(X_s)$$

$$GLRT(\mu, H_t) = \log \frac{d\mathbb{P}_\mu}{\sup_{\lambda \in alt(\mu)} d\mathbb{P}_\lambda}(H_t) = \inf_{\lambda \in alt(\mu)} \log \frac{d\mathbb{P}_\mu}{d\mathbb{P}_\lambda}(H_t)$$

Based on observations in H_t, $LRT(\mu, \lambda, H_t)$ is how likely μ is compared to λ.

$\mathbb{E}_\mu[LRT(\mu, \lambda)] = KL(H_{t,\mu}, H_{t,\lambda})$.

$GLRT(\mu, H_t)$ compares μ to its alternative set.

Stopping rule

Stopping rule: stop if $GLRT(\widehat{\mu}_t, H_t) > \log \frac{\log t}{\delta}$ (approximatively, up to constants).

Theorem 7

Any algorithm using the above stopping rule with the recommendation rule $\widehat{i} = \arg\max_k \widehat{\mu}_t^k$ verifies, for all $\mu \in \mathbb{R}^K$,

$$\mathbb{P}_\mu(\tau < +\infty \wedge \widehat{i} \neq *_\mu) \leq \delta$$

Together, stopping rule and recommendation rule ensure δ-correctness (provided that the sampling rule ensures $\tau < +\infty$).

Proof: deviation bound on $\mathbb{P}_\mu(LRT(\widehat{\mu}_t, \mu, H_t) > \varepsilon)$.

Track and Stop

While $GLRT(\widehat{\mu}, H_t) \leq \log \frac{\log t}{\delta}$,

- Compute the oracle allocation $w_{opt}(\widehat{\mu}_t)$
- If an arm has $N_t^k < \sqrt{t}$, sample it. (forced exploration)
- Otherwise, sample $k_t = \arg\min_k N_t^k - t w_{opt}^k(\widehat{\mu}_t)$ (tracking)

Recommend $\widehat{i} = \arg\max_k \widehat{\mu}_\tau^k$.

Theorem: Track and Stop is asymptotically optimal, i.e.

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_\mu[\tau]}{\log \frac{1}{\delta}} \leq \frac{1}{\sup_{w \in \triangle_K} \inf_{\lambda \in alt(\mu)} \sum_k w_k \frac{1}{2}(\mu_k - \lambda_k)^2} .$$

Issues and Improvements

Track-and-Stop is computationally intensive due to the oracle allocation computation.

Forced exploration is wasteful.

An improvement: use an iterative algorithm to compute the oracle allocation, but do only one iteration at a time.

A related improvement: use optimism in that iterative algorithm to avoid forced exploration.

$\rightarrow$ sample complexity bounds (i.e. bounds on $\mathbb{E}[\tau]$) for non-zero δ.

$\rightarrow$ still no good bound for $\delta \approx 0.1$.

Fixed Budget BAI

Fixed budget best arm identification:

- For $t = 1, \dots, T$, choose $k_t \in [K]$ and observe $X_t \sim \nu_{k_t}$.
- Recommend $\widehat{i} \in [K]$ after time T.
- Goal: minimize $\mathbb{P}(\widehat{i} \neq *)$.

Complexities : $H_1 = \sum_{k:\Delta_k>0} \frac{1}{\Delta_k^2}$, $H_2 = \max_{k:\Delta_k>0} \frac{k}{\Delta_{(k)}^2}$.

Property: $H_2 \leq H_1 \leq \log(2K) H_2$.

Lower bound: of order $\exp(-T/\log(K)H_1)$ if H_1 is unknown; of order $\exp(-T/H_1)$ if known.

UCB-E

Parameter: exploration parameter $a > 0$.
For each round $t = 1, \ldots, T$,

- Pull $k_t \in \arg\max_k \widehat{\mu}_t^k + \sqrt{\frac{a}{N_t^k}}$.

Recommend $\widehat{i} \in \arg\max_k \widehat{\mu}_T^k$.

Theorem 8

If UCB-E is run with parameter $0 < a < \frac{25}{36}\frac{T-K}{H_1}$, with $H_1 = \sum_{k:\Delta_k>0} \frac{1}{\Delta_k^2}$, then it satisfies

$$\mathbb{P}(\widehat{i} \neq *) \leq 2KT \exp\left(-\frac{2}{25}a\right) .$$

Issue: in order to match the lower bound, H_1 has to be known.

References

Thank you!

- Cesa-Bianchi, Nicolo and Gábor Lugosi. Prediction, learning, and games. Cambridge university press, 2006.
- Hazan, Elad et al. "Introduction to online convex optimization". In: Foundations and Trends® in Optimization 2.3-4 (2016), pp. 157–325.
- Lattimore, Tor and Csaba Szepesvári. Bandit algorithms. Cambridge University Press, 2020.
- Shalev-Shwartz, Shai et al. "Online learning and online convex optimization". In: Foundations and Trends® in Machine Learning 4.2 (2012), pp. 107–194.

Advertisement: formalizing probability and bandits in Lean

Lean theorem prover: https://leanprover-community.github.io/

Current state: measure theory has solid bases. Probability theory not so much. We have conditional expectation, martingales, independence.

Goal now: add results about martingales and concentration inequalities. Then we can write the proof of the regret bound of UCB.

Then we'll have machine-verified bandit proofs!

More exiting: automatic generation of proofs with machine learning. Example: gpt-f.