Mathematical approach to RL toy models

Luc Lehéricy

Laboratoire J.A. Dieudonné, CNRS, UCA

TABLE OF CONTENTS

UNIVERSITÉ CÔTE D'AZUR

1 **Regret and pseudo-regret**

2 **Stochastic bandits**

- Upper bounds: UCB algorithm
- Lower bounds

3 **"All-seeing" adversarial bandits**

4 **Adversarial bandits**

- Upper bounds: Exponential weighting algorithm
- Lower bounds

- K arms (indexed by $1 \leqslant i \leqslant K$), time horizon n
- **Random** reward (gain) of arm i at time t: $G_{i,t}$
- Strategy / forecaster / algorithm: at each time t, select an arm I_t based on the information available and an **internal randomness**, and observe the gain $G_{I_t,t}$.
- Information available at time t: choices $(I_s)_{1 \leqslant s < t}$ and gains $(G_{I_s,s})_{1 \leqslant s < t}$ observed before time t.
- Regret $R_n = \max\limits_{1 \leqslant i \leqslant K} \sum\limits_{t=1}^{n} G_{i,t} - \sum\limits_{t=1}^{n} G_{I_t,t}$ = loss compared to sticking to the best arm (which may depend on the realization of the gains)
- Pseudo-regret $\bar{R}_n = \max\limits_{1 \leqslant i \leqslant K} \mathbb{E}\left[\sum\limits_{t=1}^{n} G_{i,t}\right] - \mathbb{E}\left[\sum\limits_{t=1}^{n} G_{I_t,t}\right]$ = expected loss compared to the best expected loss when sticking to a fixed arm

We always have $\bar{R}_n \leqslant \mathbb{E}R_n$. In the following, we study the pseudo-regret $\bar{R}_n$.

Stochastic bandits

- **The arms are independent**: the sequences $(G_{i,t})_{t\geqslant 1}$, $1 \leqslant i \leqslant K$, are independent.
- **The distribution of the gains does not depend on time or on previous rewards or events**: given an arm i, the sequence $(G_{i,t})_{t\geqslant 1}$ is a sequence of i.i.d. random variables with distribution ν_i and mean μ_i.

In short: each arm is "blind" to the others and itself.

Write $\mu^\star = \max_{1\leqslant i\leqslant K} \mu_i$, $\Delta_i = \mu^\star - \mu_i$, and $N_i(t) = \sum_{s=1}^{t} \mathbf{1}_{I_s=i}$ the number of times the arm i has been selected before time t. Then

$$\bar{R}_n = n\mu^\star - \mathbb{E}\sum_{t=1}^{n} \mu_{I_t,t}$$

$$= \sum_{i=1}^{K} \Delta_i \, \mathbb{E}[N_i(n)].$$

Assume the **gains are sub-Gaussian**[1], i.e.[2] there exists $\sigma > 0$ such that for all $x > 0$, i and t,

$$\max\{\mathbb{P}(G_{i,t} < -t), \mathbb{P}(G_{i,t} > t)\} \leqslant \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Example: random variables taking values in $[0,1]$ are sub-Gaussian with $\sigma = 1/2$.

Upper Confidence Bound (α) strategy

Let $\alpha > 1$. For t from 0 to $n-1$,

- Compute the estimated gain of arm i: $\hat{\mu}_{i,t} = \frac{1}{N_i(t)} \sum_{s=1}^{t} G_{i,s} \mathbf{1}_{I_s = i}$
- Select $I_{t+1} \in \underset{1 \leqslant i \leqslant K}{\arg\max} \underbrace{\left[\hat{\mu}_{i,t} + \sqrt{\alpha \frac{2\sigma^2 \log t}{N_i(t)}}\right]}$

upper bound of a confidence interval of level $1 - t^{-\alpha}$ of μ_i

[1] An algorithm without this assumption is discussed in Bubeck & Cesa-Bianchi (2012).

[2] This is not the true definition of sub-Gaussianity, although it is equivalent up to modification of σ.

Theorem: UCB(α) pseudo-regret

$$\bar{R}_n \leqslant \log(n) \sum_{i:\Delta_i>0} \frac{8\alpha\sigma^2}{\Delta_i} + \text{cst}(\alpha) \sum_{i=1}^{K} \Delta_i$$

Let $m_{i,t} = \sqrt{\alpha \frac{2\sigma^2 \log t}{N_i(t)}}$. With probability $\geqslant 1 - 2t^{-\alpha}$, $\hat{\mu}_{i,t} \in [\mu_i - m_{i,t}, \mu_i + m_{i,t}]$.

Let $i^\star$ be a best arm and i such that $\Delta_i > 0$. Assume that

1. $m_{i,t} < \frac{\Delta_i}{2}$,
2. $\hat{\mu}_{i^\star,t} + m_{i^\star,t} \geqslant \mu^\star$ and
3. $\hat{\mu}_{i,t} \leqslant \mu_i + m_{i,t}$.

Then $\hat{\mu}_{i,t} + m_{i,t} < \mu^\star \leqslant \hat{\mu}_{i^\star,t} + m_{i^\star,t}$, so $I_t \neq i$. Thus, $I_t = i$ is possible only when at least one of the above is false.

Theorem: UCB(α) pseudo-regret

$$\bar{R}_n \leqslant \log(n) \sum_{i:\Delta_i>0} \frac{8\alpha\sigma^2}{\Delta_i} + \text{cst}(\alpha) \sum_{i=1}^{K} \Delta_i$$

Let $m_{i,t} = \sqrt{\alpha \frac{2\sigma^2 \log t}{N_i(t)}}$. With probability $\geqslant 1 - 2t^{-\alpha}$, $\hat{\mu}_{i,t} \in [\mu_i - m_{i,t}, \mu_i + m_{i,t}]$.

Let $i^\star$ be a best arm and i such that $\Delta_i > 0$. Assume that

1. $m_{i,t} < \frac{\Delta_i}{2}$, $\quad \Leftarrow N_i(t) > \frac{8\alpha\sigma^2 \log t}{\Delta_i^2}$
2. $\hat{\mu}_{i^\star,t} + m_{i^\star,t} \geqslant \mu^\star$ and $\quad$ (happens with proba $\geqslant 1 - t^{-\alpha}$)
3. $\hat{\mu}_{i,t} \leqslant \mu_i + m_{i,t}$. $\quad$ (happens with proba $\geqslant 1 - t^{-\alpha}$)

Then $\hat{\mu}_{i,t} + m_{i,t} < \mu^\star \leqslant \hat{\mu}_{i^\star,t} + m_{i^\star,t}$, so $I_t \neq i$. Thus, $I_t = i$ is possible only when at least one of the above is false.

Theorem: UCB(α) pseudo-regret

$$\bar{R}_n \leqslant \log(n) \sum_{i:\Delta_i>0} \frac{8\alpha\sigma^2}{\Delta_i} + \mathrm{cst}(\alpha) \sum_{i=1}^{K} \Delta_i$$

Let $m_{i,t} = \sqrt{\alpha \frac{2\sigma^2 \log t}{N_i(t)}}$. With probability $\geqslant 1 - 2t^{-\alpha}$, $\hat{\mu}_{i,t} \in [\mu_i - m_{i,t}, \mu_i + m_{i,t}]$.

Let $i^\star$ be a best arm and i such that $\Delta_i > 0$. Assume that

1. $m_{i,t} < \frac{\Delta_i}{2}$, $\quad \Leftarrow N_i(t) > \frac{8\alpha\sigma^2 \log t}{\Delta_i^2}$
2. $\hat{\mu}_{i^\star,t} + m_{i^\star,t} \geqslant \mu^\star$ and $\quad$ (happens with proba $\geqslant 1 - t^{-\alpha}$)
3. $\hat{\mu}_{i,t} \leqslant \mu_i + m_{i,t}$. $\quad$ (happens with proba $\geqslant 1 - t^{-\alpha}$)

Then $\hat{\mu}_{i,t} + m_{i,t} < \mu^\star \leqslant \hat{\mu}_{i^\star,t} + m_{i^\star,t}$, so $I_t \neq i$. Thus, $I_t = i$ is possible only when at least one of the above is false, hence

$$\mathbb{E}[N_i(n)] \leqslant \frac{8\alpha\sigma^2 \log n}{\Delta_i^2} + \mathbb{E}\sum_{t=1}^{n} \mathbf{1}_{I_t=i}\mathbf{1}_{\text{1. true}} \leqslant \frac{8\alpha\sigma^2 \log n}{\Delta_i^2} + \sum_{t=1}^{n} 2t^{-\alpha}.$$

and finally use $\bar{R}_n = \sum_{i=1}^{K} \Delta_i \, \mathbb{E}[N_i(n)]$.

Let $\mathcal{D}$ be a set of probability measures on $\mathbb{R}$. Any element $\nu = (\nu_1, \dots, \nu_K) \in \mathcal{D}^K$ is identified with the bandit problem with arm distributions given by ν, and the corresponding probability measures and expectation are denoted $\mathbb{P}_\nu$ and $\mathbb{E}_\nu$. We write $\mu(\nu_i) = \mathbb{E}_{X\sim\nu_i}[X]$ the expected gain of arm i under ν.

A strategy is uniformly fast convergent on $\mathcal{D}$ if for all $\nu \in \mathcal{D}^K$, for all sub-optimal arms i w.r.t. ν and for all $\alpha \in (0, 1]$, $\mathbb{E}_\nu[N_i(t)] = o(t^\alpha)$.

In the following, let $\nu = (\nu_1, \dots, \nu_K)$ be a K-uple of probability measures and assume the strategy is uniformly fast convergent on a set $\mathcal{D}$.

Theorem: Lower bound for the pseudo-regret

For any sub-optimal arms i and any measures $\nu_i' \in \mathcal{D}$ such that $\mu(\nu_i') > \max_j \mu(\nu_j)$,

$$\liminf_{n\to+\infty} \frac{\mathbb{E}_\nu[N_i(n)]}{\log(n)} \geqslant \frac{1}{\mathrm{KL}(\nu_i, \nu_i')}$$

where the Kullback Leibler divergence defined by $\mathrm{KL}(\nu_i, \nu_i') = \mathbb{E}_{\nu_i}[\log \frac{\mathrm{d}\nu_i}{\mathrm{d}\nu_i'}]$ is always nonnegative and equals zero if and only if $\nu_i = \nu_i'$.

When X is a random variable, write $\mathbb{P}_\nu^X$ the distribution of X under $\mathbb{P}_\nu$.

Let $\mathbf{X}_t = (I_1, G_{I_1,1}, \dots, I_t, G_{I_t,t})$ be the sequence of arms selected and gains observed.

Fundamental inequality (Garivier et al., 2019)

For all $\nu, \nu' \in \mathcal{D}$, for all $\sigma(\mathbf{X}_t)$-measurable random variable Z taking values in $[0,1]$,

$$\sum_{i=1}^{K} \mathbb{E}_\nu[N_i(t)]\mathrm{KL}(\nu_i, \nu_i') = \mathrm{KL}(\mathbb{P}_\nu^{\mathbf{X}_t}, \mathbb{P}_{\nu'}^{\mathbf{X}_t}) \geqslant kl(\mathbb{E}_\nu[Z], \mathbb{E}_{\nu'}[Z])$$

where $kl(p,q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$.

The right-most inequality (admitted) holds for any distribution. $Z \in [0,1]$ matters!

Informally, the KL divergence is how different the distributions look. Here the only visible difference is through the pulled armed: at time s, the difference "increases" by how different the distributions of the current arm are: $\mathrm{KL}(\nu_{I_s}, \nu'_{I_s})$.

Then $\mathbb{E}_\nu\left[\sum_{s=1}^t \mathrm{KL}(\nu_{I_s}, \nu'_{I_s})\right] = \sum_{i=1}^K \mathbb{E}_\nu[N_i(t)]\mathrm{KL}(\nu_i, \nu_i')$.

Chain rule:

$$\mathrm{KL}(\mathbb{P}_\nu^{\mathbf{X}_t}, \mathbb{P}_{\nu'}^{\mathbf{X}_t})$$

$$= \mathbb{E}_\nu[\mathrm{KL}(\mathbb{P}_\nu^{I_t, G_{I_t,t}|\mathbf{X}_{t-1}}, \mathbb{P}_{\nu'}^{I_t, G_{I_t,t}|\mathbf{X}_{t-1}})] + \mathrm{KL}(\mathbb{P}_\nu^{\mathbf{X}_{t-1}}, \mathbb{P}_{\nu'}^{\mathbf{X}_{t-1}})$$

$$= \mathbb{E}_\nu \left[\sum_{i=1}^K \mathbb{P}_\nu(I_t = i|\mathbf{X}_t) \int_g p_\nu(G_{I_t,t} = g|I_t = i, \mathbf{X}_t) \log \frac{\mathbb{P}_\nu(I_t = i|\mathbf{X}_t) p_\nu(G_{I_t,t} = g|I_t = i, \mathbf{X}_t)}{\mathbb{P}_{\nu'}(I_t = i|\mathbf{X}_t) p_{\nu'}(G_{I_t,t} = g|I_t = i, \mathbf{X}_t)} \right.$$

$$\left. + \mathrm{KL}(\mathbb{P}_\nu^{\mathbf{X}_{t-1}}, \mathbb{P}_{\nu'}^{\mathbf{X}_{t-1}}) \right]$$

$$= \mathbb{E}_\nu \left[\sum_{i=1}^K \mathbb{P}_\nu(I_t = i|\mathbf{X}_t) \int_g p_\nu(G_{i,t} = g) \log \frac{p_\nu(G_{i,t} = g)}{p_{\nu'}(G_{i,t} = g)} \right] + \mathrm{KL}(\mathbb{P}_\nu^{\mathbf{X}_{t-1}}, \mathbb{P}_{\nu'}^{\mathbf{X}_{t-1}})$$

since the algorithm choosing I_t only takes the past $\mathbf{X}_t$ into account, not ν: $\mathbb{P}_\nu(I_t = i|\mathbf{X}_t) = \mathbb{P}_{\nu'}(I_t = i|\mathbf{X}_t)$, and the gains of an arm are independent from the past.

$$(\ldots) = \mathbb{E}_\nu \left[\sum_{i=1}^K \mathbf{1}_{I_t=i} \mathrm{KL}(\nu_i, \nu_i') \right] + \mathrm{KL}(\mathbb{P}_\nu^{\mathbf{X}_{t-1}}, \mathbb{P}_{\nu'}^{\mathbf{X}_{t-1}})$$

$$= \mathbb{E}_\nu \left[\sum_{i=1}^K N_i(t) \right] \mathrm{KL}(\nu_i, \nu_i').$$

Fundamental inequality (Garivier et al., 2019)

For all $\nu, \nu' \in \mathcal{D}^K$, for all $\sigma(\mathbf{X}_t)$-measurable random variable Z taking values in $[0,1]$,

$$\sum_{i=1}^{K} \mathbb{E}_\nu[N_i(t)]\mathrm{KL}(\nu_i, \nu_i') \geqslant kl(\mathbb{E}_\nu[Z], \mathbb{E}_{\nu'}[Z]).$$

Take $Z = N_i(t)/t$, this is indeed $\sigma(\mathbf{X}_t)$-measurable and in $[0,1]$.

Also use $kl(p,q) = \underbrace{p \log \frac{1}{q}}_{\geqslant 0} + (1-p)\log\frac{1}{1-q} + \underbrace{p\log p + (1-p)\log(1-p)}_{= kl(p,1/2) - \log 2 \geqslant -\log 2}$,

so that $$kl(\mathbb{E}_\nu[Z], \mathbb{E}_{\nu'}[Z]) \geqslant \left(1 - \frac{\mathbb{E}_\nu[N_i(t)]}{t}\right) \log \frac{t}{t - \mathbb{E}_{\nu'}[N_i(t)]} - \log 2.$$

Let i be a non-optimal arm. Take $\nu' = (\dots, \nu_{i-1}, \nu_i', \nu_{i+1}, \dots)$ where $\mu(\nu_i') > \mu^\star$.

Uniformly fast convergent: $\forall \alpha > 0, \mathbb{E}_\nu[N_i(t)] = o(t^\alpha)$ and $\mathbb{E}_{\nu'}[N_i(t)] = t - o(t^\alpha)$:

$$\mathbb{E}_\nu[N_i(t)]\mathrm{KL}(\nu_i, \nu_i') \geqslant kl(\mathbb{E}_\nu[Z], \mathbb{E}_{\nu'}[Z]) \geqslant (1 - o(1)) \log \frac{t}{o(t^\alpha)} - \log 2 \quad \Rightarrow \log t.$$

Theorem: UCB(α) pseudo-regret

When the gains are σ-sub-Gaussian,

$$\limsup \frac{\bar{R}_n}{\log(n)} \leqslant \sum_{i:\Delta_i>0} \frac{8\alpha\sigma^2}{\Delta_i}.$$

Theorem: Lower bound for the pseudo-regret

For any strategy,

$$\liminf_{n\to+\infty} \frac{\bar{R}_n}{\log(n)} \geqslant \sum_{i:\Delta_i>0} \frac{\Delta_i}{\inf_{\nu_i':\mu(\nu_i')>\max_j \mu_j} \mathrm{KL}(\nu_i,\nu_i')}.$$

The bounds match (pseudo-regret $\propto \log(n)$) up to a constant that can be very large (take Bernoulli r.v. with parameters tending to 0 or 1). Variants of UCB may improve the upper bound in other situations.

"All-seeing" adversarial bandits

- **The environment is a player**: at each time step, the environment decides what the new gains are depending on "the past" (see third point) and an internal randomness. In particular, **no "nice" assumption can be made on the environment**: the sequences $(G_{i,t})_{t\geqslant 1}$, $1 \leqslant i \leqslant K$, are no longer independent, the distribution of the gains may depend on time.
- **The environment and the forecaster select an arm and choose the gains at the same time**: the environment may not use the current move of the forecaster to decide the gains.
- **The environment may adapt to the past moves of the forecaster**: $(G_{i,t})_{1\leqslant i\leqslant K}$ may depend on $(I_s)_{s<t}$, in addition to $(G_{i,s})_{1\leqslant i\leqslant K, s<t}$.

$\Rightarrow$ Two competing players. Goal: minimize the pseudo-regret without knowledge of the environment, that is minimize

$$\sup_{\text{environment}} \bar{R}_n = \sup_{\text{environment}} \left\{ \max_{1\leqslant i\leqslant K} \mathbb{E}\left[\sum_{t=1}^{n} G_{i,t}\right] - \mathbb{E}\left[\sum_{t=1}^{n} G_{I_t,t}\right] \right\}.$$

Is it possible to have sub-linear pseudo-regret?

"ALL-SEEING" ADVERSARIAL BANDITS

UNIVERSITÉ CÔTE D'AZUR

When minimizing $\sup_{\text{environment}} \bar{R}_n$, consider the worst possible environment given our strategy: the environment knows the forecaster's strategy.

Given the environment also knows the past moves of the players as well as the gains, it has as much information as the player and hence knows the probability distribution of the next selected arm.

Consider the following environment:

- Start by giving gain $g \in (0, 1)$ to all arms as long as the forecaster has not switched arms since the start.
- If at some point the forecaster had a probability non-zero to switch arm, keep giving gain g to the arm with lowest probability of being selected and 0 to the other arms.

When minimizing $\sup_{\text{environment}} \bar{R}_n$, consider the worst possible environment given our strategy: the environment knows the forecaster's strategy.

Given the environment also knows the past moves of the players as well as the gains, it has as much information as the player and hence knows the probability distribution of the next selected arm.

Consider the following environment:

- Start by giving gain $g \in (0, 1)$ to all arms as long as the forecaster has not switched arms since the start.
- If at some point the forecaster had a probability non-zero to switch arm, keep giving gain g to the arm with lowest probability of being selected and 0 to the other arms.

Result: if the strategy is non-constant, the average gain per time step after the first switch will always be no larger than $g/2$, compared to g if the strategy was constant: **linear pseudo-regret**.

If the strategy is constant, then an environment that rewards g to the selected arm and 1 to a different arm has **linear pseudo-regret**.

When minimizing $\sup_{\text{environment}} \bar{R}_n$, consider the worst possible environment given our strategy: the environment knows the forecaster's strategy.

Given the environment also knows the past moves of the players as well as the gains, it has as much information as the player and hence knows the probability distribution of the next selected arm.

Consider the following environment:

- Start by giving gain $g \in (0, 1)$ to all arms as long as the forecaster has not switched arms since the start.
- If at some point the forecaster had a probability non-zero to switch arm, keep giving gain g to the arm with lowest probability of being selected and 0 to the other arms.

Result: if the strategy is non-constant, the average gain per time step after the first switch will always be no larger than $g/2$, compared to g if the strategy was constant: **linear pseudo-regret**.

If the strategy is constant, then an environment that rewards g to the selected arm and 1 to a different arm has **linear pseudo-regret**.

Adversarial bandits

- **The environment is a player**: at each time step, the environment decides what the new gains are depending on "the past" (see third point) and an internal randomness. In particular, **no "nice" assumption can be made on the environment**: the sequences $(G_{i,t})_{t\geqslant 1}$, $1 \leqslant i \leqslant K$, are no longer independent, the distribution of the gains may depend on time.
- **The environment and the forecaster select an arm and choose the gains at the same time**: the environment may not use the current move of the forecaster to decide the gains.
- **The environment can not see the moves of the forecaster**: conditionally to $(G_{i,s})_{1\leqslant i\leqslant K, s<t}$, $(G_{i,t})_{1\leqslant i\leqslant K}$ is independent of $(I_s)_{s\leqslant t}$.

Similar to **expert aggregation**, in which the forecaster observes all gains after having chosen instead of just the chosen one.

Assume the rewards take value in $[0,1]$, and define the losses as $\ell_{i,t} = 1 - G_{i,t}$.

Exponential weight for Exploration and Exploitation strategy (Exp3)

Let $(\eta_t)_{t\geqslant 1}$ be a **nonincreasing** sequence of **positive** numbers.
Let p_1 be the uniform distribution over $\{1,\dots,K\}$ and $\tilde{L}_{1,0} = 0$ for all $1 \leqslant i \leqslant K$.
For each $t = 1,\dots,n$,

1. Draw $I_t \sim p_t$,
2. For each i, update the estimated cumulated loss of arm i:

$$\tilde{L}_{i,t} = \tilde{L}_{i,t-1} + \underbrace{\frac{\ell_{i,t}}{p_{i,t}}\mathbf{1}_{I_t=i}}_{\text{estimated loss } \tilde{\ell}_{i,t}\,:\,\mathbb{E}[\tilde{\ell}_{i,t}]=\ell_{i,t}}$$

3. Update the probability distribution: $p_{i,t+1} = \dfrac{\exp(-\eta_{t+1}\tilde{L}_{i,t})}{\sum_k \exp(-\eta_{t+1}\tilde{L}_{k,t})}$

Results from Bubeck et al., Section 3.

Theorem: Exp3 upper bounds

When running Exp3 with...

$$\eta_t = \sqrt{\frac{2\log K}{nK}} \quad \text{(known time horizon), pseudo regret} \quad \bar{R}_n \leqslant \sqrt{2nK\log K},$$

$$\eta_t = \sqrt{\frac{\log K}{tK}} \quad \text{(unknown time horizon), pseudo regret} \quad \bar{R}_n \leqslant 2\sqrt{nK\log K}.$$

Holds for any environment!

High probability bounds for the **regret** are also available for a variant of Exp3, of the form: with probability at least $1-\delta$,

$$R_n \leqslant 5.15\sqrt{nK\log K} + \sqrt{\frac{nK}{\log K}}\log\frac{1}{\delta},$$

and

$$\bar{R}_n \leqslant \mathbb{E}R_n \leqslant 5.15\sqrt{nK\log K} + \sqrt{\frac{nK}{\log K}}.$$

Let $\epsilon > 0$. Consider a stochastic bandit with reward distributions ν such that $\nu_1 = \mathcal{B}(\frac{1}{2} + \varepsilon)$ and $\nu_i = \mathcal{B}(\frac{1}{2})$ for $i \neq 1$.

In order to distinguish a $\mathcal{B}(\frac{1}{2})$ and a $\mathcal{B}(\frac{1}{2} + \varepsilon)$, the forecaster needs a number N of observations that satisfies $\mathrm{KL}(\mathcal{B}(\frac{1}{2}), \mathcal{B}(\frac{1}{2} + \varepsilon)) \geqslant 1/N$, that is $\varepsilon^2 \geqslant 1/N$ when ε is small.

Since before identifying the best arm the strategy is essentially random, $N = n/K$, so that $\varepsilon = \sqrt{K/n}$. The pseudo-regret is then of order

$$(n - \underbrace{n/K}_{\text{number of times the correct arm is selected}}) \times \varepsilon \simeq \sqrt{nK}$$

Theorem: minimax lower bound

$$\inf_{\text{strategy}} \sup_{\substack{\text{stochastic bandit} \\ \text{with Bernoulli rewards}}} \bar{R}_n \geqslant \frac{1}{20}\sqrt{nK}.$$

Matches the upper bounds up to a $\sqrt{\log K}$ factor.

REFERENCES

Bubeck, S., & Cesa-Bianchi, N. (2012). **Regret analysis of stochastic and nonstochastic multi-armed bandit problems.** arXiv preprint arXiv:1204.5721.

Garivier, A., Ménard, P., & Stoltz, G. (2019). **Explore first, exploit next: The true shape of regret in bandit problems.** Mathematics of Operations Research, 44(2), 377-399.

Thank you!