STATS 701 – Theory of Reinforcement Learning
Bandit Problems

Ambuj Tewari

Associate Professor, Department of Statistics, University of Michigan
tewaria@umich.edu
https://ambujtewari.github.io/stats701-winter2021/

Slide Credits: Alan Malek @ DeepMind

Winter 2021

The Basic Bandit Game

Protocol: Finite-Arm Bandits

Given: game length T, number of arms K

For $t = 1, 2, \ldots, T$,

- The learner picks action $I_t \in \{1, \ldots, K\}$
- The adversary simultaneously picks rewards $r_t \in \{1, \ldots, K\} \to [0, 1]$
- The learner observes and receives $r_t(I_t)$
- The learner does not observe $r_t(i)$ for $i \neq I_t$

The goal: control the regret (a random variable)

$$\mathcal{R}_T = \underbrace{\max_i \sum_{t=1}^{T} r_t(i)}_{\text{Best action in hindsight}} - \sum_{t=1}^{T} r_t(I_t)$$

Bandits are Super Simple MDP

- $S = \{\text{the\_state}\}$, $P(\text{the\_state}|\text{the\_state}, a) = 1$
- Why should we care about this in RL?
  - Creates a tension between
    - Exploration (learning about the loss of actions)
    - Exploitation (playing actions that will have low regret)
  - Exploration/Exploitation is absent in full-information but very present in reinforcement learning
  - Model is simple enough to allow for comprehensive theory
  - Easily incorporates adversarial data
  - Useful algorithm design principles

The Regret

$$\mathcal{R}_T = \underbrace{\max_i \sum_{t=1}^{T} r_t(i)}_{\text{Best action in hindsight}} - \sum_{t=1}^{T} r_t(I_t)$$

- $\mathcal{R}_T$ is a random variable we do not observe
- Different objectives, from easiest to hardest
  - Pseudo-regret $\overline{\mathcal{R}_T} = \max_i \mathbb{E}\left[\sum_{t=1}^{T} r_t(i)\right] - \mathbb{E}\left[\sum_{t=1}^{T} r_t(I_t)\right]$
  - Expected regret $\mathbb{E}[\mathcal{R}_T] = \mathbb{E}\left[\underbrace{\max_i \sum_{t=1}^{T} r_t(i)}_{\text{can depend on } I_t} - \sum_{t=1}^{T} r_t(I_t)\right]$
  - High probability bounds on the realized regret

The Regret

$$\mathcal{R}_T = \underbrace{\max_i \sum_{t=1}^{T} r_t(i)}_{\text{Best action in hindsight}} - \sum_{t=1}^{T} r_t(I_t)$$

- $\mathcal{R}_T$ is a random variable we do not observe
- Different objectives, from easiest to hardest
  - Pseudo-regret $\overline{\mathcal{R}_T} = \max_i \mathbb{E}\left[\sum_{t=1}^{T} r_t(i)\right] - \mathbb{E}\left[\sum_{t=1}^{T} r_t(I_t)\right]$
  - Expected regret $\mathbb{E}[\mathcal{R}_T] = \mathbb{E}\left[\underbrace{\max_i \sum_{t=1}^{T} r_t(i)}_{\text{can depend on } I_t} - \sum_{t=1}^{T} r_t(I_t)\right]$
  - High probability bounds on the realized regret
- We always have $\overline{\mathcal{R}_T} \le \mathbb{E}[\mathcal{R}_T]$
- If the adversary is *reactive*, then the distribution of r_t can be a function of $I_1, \ldots, I_{t-1}$
- Otherwise, the adversary is *oblivious* and $\overline{\mathcal{R}_T} = \mathbb{E}[\mathcal{R}_T]$

Our Focus

- Introduce most popular bandit problems
  - Adversarial Bandits
  - Stochastic Bandits
  - Pure Exploration Bandits
  - Contextual Bandits (time permitting)
- Concentrate on useful algorithm design principles
  - Exponential weights (still useful)
  - Optimism in the face of Uncertainty
  - Probability matching (i.e. Thompson sampling)
  - Action-Elimination

Other Settings that Have Been Considered

- Data models for r_t
  - chosen by an adversary
  - sampled i.i.d.
  - stochastic with adversarial perturbations...
- Action spaces
  - Finite number of arms
  - A vector space (r_t are functions)
  - Combinatorial (e.g. subsets, paths on a graph)
- Objectives
  - Pseudo-regret (the expectation over the learner's randomness)
  - Realized regret (with high probability)
  - Best-arm identification a.k.a. pure exploration
- Side information
  - Linear rewards
  - Competing with a policy class
- ...

Adversarial Protocol

Protocol: Finite-Arm Adversarial Bandits

Given: game length T, number of arms K

For $t = 1, 2, \dots, T$,

- The learner picks action $I_t \in \{1, \dots, K\}$
- The adversary simultaneously picks losses $\ell_t \in [0, 1]^K$
- The learner observes and receives $\ell_t(I_t)$

- The results are easier to state using losses instead of rewards
- Randomization of I_t is *essential*
- We are familiar with adversarial data from the first half
- The simple idea of estimating ℓ_t from $\ell_t(I_t)$ and then applying a full-information algorithm works very well

Algorithm Design Principle: Exponential Weights

Algorithm: Exp3 [Auer et al., 2002b]

Given: number of arms K, learning rate $\eta > 0$, length T
Initialize $p_1(i) = 1/K$, $\hat{L}_0(i) = 0$ for all $i \in [K]$

For $t = 1, 2, \ldots, T$:

- Sample $I_t \sim p_t$ and observe $\ell_t(I_t)$
- Estimate $\hat{\ell}_t(i) = \frac{\ell_t(I_t)}{p_t(I_t)} \mathbb{1}_{\{I_t = i\}}$ and $\hat{L}_t = \hat{\ell}_t + \hat{L}_{t-1}$
- Calculate $W_t = \sum_j e^{-\eta \hat{L}_t(j)}$ and $p_{t+1}(i) = \frac{1}{W_t} e^{-\eta \hat{L}_t(i)}$

Algorithm Design Principle: Exponential Weights

Algorithm: Exp3 [Auer et al., 2002b]

Given: number of arms K, learning rate $\eta > 0$, length T
Initialize $p_1(i) = 1/K$, $\hat{L}_0(i) = 0$ for all $i \in [K]$

For $t = 1, 2, \ldots, T$:

- Sample $I_t \sim p_t$ and observe $\ell_t(I_t)$
- Estimate $\hat{\ell}_t(i) = \frac{\ell_t(I_t)}{p_t(I_t)} \mathbb{1}_{\{I_t = i\}}$ and $\hat{L}_t = \hat{\ell}_t + \hat{L}_{t-1}$
- Calculate $W_t = \sum_j e^{-\eta \hat{L}_t(j)}$ and $p_{t+1}(i) = \frac{1}{W_t} e^{-\eta \hat{L}_t(i)}$

- Exp3 = Exponential Weights for Exploration and Exploitation
- $\hat{\ell}_t$ is the importance-weighted estimator of ℓ_t
- $\hat{\ell}_t$ is unbiased:

$$\mathbb{E}_{I_t \sim p_t}[\hat{\ell}_t(i)] = \mathbb{E}\left[\frac{\ell_t(I_t)}{p_t(I_t)} \mathbb{1}_{\{I_t = i\}}\right] = \sum_j p_t(j) \frac{\ell_t(j)}{p_t(j)} \mathbb{1}_{\{j = i\}} = \ell_t(i).$$

- Exp3 runs exponential weights on $\hat{\ell}_t$

Exp3: Analysis

$$
\begin{aligned}
\frac{W_t}{W_{t-1}} &= \sum_{j=1}^{K} \frac{w_t(j)}{W_{t-1}} \\
&= \sum_j \frac{w_{t-1}(j)}{W_{t-1}} \exp(-\eta \hat{\ell}_t(j)) \\
&= \sum_j p_t(j) \exp(-\eta \hat{\ell}_t(j)) \\
&\le \sum_j p_t(j) \left(1 - \eta \hat{\ell}_t(j) + \frac{1}{2}\eta^2 \hat{\ell}_t(i)^2\right) && e^{-x} \le 1 - x + \frac{x^2}{2}, \forall x \ge 0 \\
&= 1 - \eta \sum_j p_t(j)\hat{\ell}_t(j) + \frac{\eta^2}{2} \sum_i p_t(j)\hat{\ell}_t(j)^2 \\
&\le \exp\left(-\eta \sum_j p_t(j)\hat{\ell}_t(j) + \frac{\eta^2}{2} \sum_j p_t(j)\hat{\ell}_t(j)^2\right) && 1 + x \le e^x, \forall x
\end{aligned}
$$

Exp3: Analysis

- Following the EW analysis, W_t is a potential function
- For any i^*, $e^{-\eta \hat{L}_T(i^*)} \leq \sum_j e^{-\eta \hat{L}_T(j)} = W_T = W_0 \prod_{t=1}^T \frac{W_t}{W_{t-1}}$.
- Bound W_0 by K and the product using the bound on the previous slide

Therefore, we have

$$e^{-\eta \hat{L}_T(i^*)} \leq W_0 \prod_{t=1}^{T} \frac{W_t}{W_{t-1}} \leq K \prod_{t=1}^{T} e^{-\eta \sum_j p_t(j)\hat{\ell}_t(j) + \frac{\eta^2}{2} \sum_j p_t(j)\hat{\ell}_t(j)^2}$$

Exp3: Analysis

$$e^{-\eta \hat{L}_T(i^*)} \le W_0 \prod_{t=1}^{T} \frac{W_t}{W_{t-1}} \le K \prod_{t=1}^{T} e^{-\eta \sum_j p_t(j)\hat{\ell}_t(j) + \frac{\eta^2}{2} \sum_j p_t(j)\hat{\ell}_t(j)^2}$$

$$\Leftrightarrow -\eta \hat{L}_T(i^*) \le \log(K) - \eta \sum_t \sum_j p_t(j)\hat{\ell}_t(j) + \frac{\eta^2}{2} \sum_t \sum_j p_t(j)\hat{\ell}_t(j)^2$$

$$\Leftrightarrow \sum_{t=1}^{T} \sum_j p_t(j)\hat{\ell}_t(j) - \hat{L}_T(i^*) \le \frac{\log(K)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_j p_t(j)\hat{\ell}_t(j)^2$$

$$\Rightarrow \sum_{t=1}^{T} \sum_j p_t(j)\mathbb{E}[\hat{\ell}_t(j)] - \mathbb{E}[\hat{L}_T(i^*)] \le \frac{\log(K)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \mathbb{E}\left[\sum_j p_t(j)\hat{\ell}_t(j)^2\right]$$

$$\Leftrightarrow \sum_{t=1}^{T} \mathbb{E}[\ell(I_t)] - L_T(i^*) \le \frac{\log(K)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \underbrace{\mathbb{E}\left[\sum_j p_t(j) \frac{\ell_t(I_t)^2}{p_t(I_t)^2} \mathbb{1}_{\{I_t = i\}}\right]}_{\text{variance term}}$$

Exp3: Analysis

Bounding the variance term turns out to be easy:

$$\mathbb{E}\left[\sum_j p_t(j)\frac{\ell_t(I_t)^2}{p_t(I_t)^2}\mathbb{1}_{\{I_t=i\}}\right] \le \mathbb{E}\left[\sum_j p_t(j)\frac{\mathbb{1}_{\{I_t=i\}}}{p_t(I_t)^2}\right]$$
$$= \mathbb{E}\left[\frac{1}{p_t(I_t)}\right] = K$$

So, plugging this in, $\sum_{t=1}^T \mathbb{E}[\ell(I_t)] - L_T(i^*) \le \frac{\log(K)}{\eta} + \frac{\eta}{2}TK$

Theorem (Exp3 upper bound [Auer et al., 2002b])

With $\eta = \sqrt{\frac{2\log(T)}{TK}}$, Exp3 has $\overline{\mathcal{R}}_T \le \sqrt{2TK\log(K)}$.

Only get pseudo-Regret bounds because the i^* in the proof was fixed, not a function of $I_1, \ldots, I_T$

Lower Bounds

Theorem (Adversarial Bandits lower bound [Auer et al., 2002b])

Any adversarial bandit algorithm must have

$$\overline{\mathcal{R}}_T = \Omega(\sqrt{TK})$$

- Exp3 upper bound: $\overline{\mathcal{R}}_T \leq \sqrt{2TK \log(K)}$
- First matching upper bound achieved by INF [Audibert and Bubeck, 2009] (which is Mirror Descent)

Upgrades

- High Probability bounds: requires a lower-variance estimate of $\hat{\ell}_t$ or an algorithm that keeps $p_t(i)$ away from zero
  - Exp3.P [Auer et al., 2002b] uses $\tilde{p}_t(i) = \gamma p_t(i) + (1-\gamma)/K$
  - Exp3-IX [Neu, 2015] uses $\hat{\ell}_t(i) = \frac{\mathbb{1}_{\{I_t=i\}}\ell_t(I_t)}{p_t(I_t)+\gamma}$
- Experts with bandits; each arm is an expert that recommends actions: you compete with the best expert (Exp4 algorithm) [Auer et al., 2002b]
- Competing with strategies that can switch [Auer, 2002]
- Feedback determined by a graph [Mannor and Shamir, 2011]
- Partial Monitoring [Bartók et al., 2014]
- Combinatorial action spaces...

Protocol

Protocol: Stochastic Bandits

Given: game length T, number of arms K, reward distributions $\nu_1, \ldots, \nu_K$

For $t = 1, 2, \ldots, T$,

- The learner picks action $I_t \in \{1, \ldots, K\}$
- The learner observes and receives reward $X_t \sim \nu_{I_t}$

Protocol

Protocol: Stochastic Bandits

Given: game length T, number of arms K, reward distributions $\nu_1, \ldots, \nu_K$

For $t = 1, 2, \ldots, T$,

- The learner picks action $I_t \in \{1, \ldots, K\}$
- The learner observes and receives reward $X_t \sim \nu_{I_t}$

- Stochastic bandits is an old problem [Thompson, 1933]
- We will use the following notation
  - Reward of arm i is sampled from ν_i with $\mu_i := \mathbb{E}_{X \sim \nu_i}[X]$
  - $i^* = \arg\max_i \mu_i$ is the best arm
  - Gaps $\Delta_i := \mu_{i^*} - \mu_i \geq 0$,
  - Number of pulls $N_{i,t} := \sum_{s=1}^{t} \mathbb{1}_{\{I_s = i\}}$
  - Empirical mean $\hat{\mu}_{i,t} := \frac{\sum_{s=1}^{t} X_s \mathbb{1}_{\{I_s = i\}}}{N_{i,t}}$

Protocol

Protocol: Stochastic Bandits

Given: game length T, number of arms K, reward distributions $\nu_1, \ldots, \nu_K$

For $t = 1, 2, \ldots, T$,

- The learner picks action $I_t \in \{1, \ldots, K\}$
- The learner observes and receives reward $X_t \sim \nu_{I_t}$

- We still want to minimize the expected regret, which has the useful decomposition

$$\mathbb{E}[\mathcal{R}_T] = T\mu_{i^*} - \sum_{t=1}^{T} \mathbb{E}[X_t] = \sum_i \Delta_i \mathbb{E}[N_{i,T}]$$

Protocol

Protocol: Stochastic Bandits

Given: game length T, number of arms K, reward distributions $\nu_1, \ldots, \nu_K$

For $t = 1, 2, \ldots, T$,

- The learner picks action $I_t \in \{1, \ldots, K\}$
- The learner observes and receives reward $X_t \sim \nu_{I_t}$

- We still want to minimize the expected regret, which has the useful decomposition

$$\mathbb{E}[\mathcal{R}_T] = T\mu_{i^*} - \sum_{t=1}^T \mathbb{E}[X_t] = \sum_i \Delta_i \mathbb{E}[N_{i,T}]$$

Assumption: 1-sub-Gaussian reward distributions

For all stochastic bandit problems, we will assume that all arms are 1-sub-Gaussian, i.e. $\mathbb{E}_{X\sim\mu}[e^{\lambda(X-\mu)^2-\lambda^2/2}] \leq 1$. For $X_1, \ldots, X_t$, This implies the Hoeffding bound

$$P\left(\frac{1}{t}\sum_{s=1}^{t} X_s - \mu_i \geq \epsilon\right) \leq e^{-\frac{\epsilon^2 t}{2}}.$$

Warm-up: Explore-Then-Commit

Algorithm: Explore-Then-Commit

Given: Game length T, exploration parameter M
For $t = 1, 2, \ldots, MK$:

- Choose $i_t = (t \mod K)$, see $X_t \sim \nu_{i_t}$

Compute empirical means $\hat{\mu}_{i,mK}$
For $t = MK + 1, MK + 2, \ldots, T$:

- Pull arm $i = \arg\max_i \hat{\mu}_{i,mK}$

- The first strategy you might try
- A proof idea that we will return to: bound regret by first bounding $\mathbb{E}[N_{i,T}]$.
- In this simple algorithm,

$$\mathbb{E}[N_{i,T}] = M + (T - MK)P\left(i = \arg\max_j \hat{\mu}_{j,MK}\right)$$

Explore-Then-Commit Upper Bound

Using the sub-Gaussian concentration bound,

$$\begin{aligned} P\left(i = \arg\max_j \hat{\mu}_{j,MK}\right) &\le P\left(\hat{\mu}_{i,MK} \ge \hat{\mu}_{i^*,MK}\right) \\ &= P\left((\hat{\mu}_{i,MK} - \mu_i) \ge (\hat{\mu}_{i^*,MK} - \mu_{i^*}) + \Delta_i\right) \\ &\le e^{-\frac{M\Delta_i^2}{4}} \text{ (the difference is } \sqrt{2/M}\text{-sub-Gaussian)} \end{aligned}$$

Explore-Then-Commit Upper Bound

Using the sub-Gaussian concentration bound,

$$P\left(i = \arg\max_j \hat{\mu}_{j,MK}\right) \leq P\left(\hat{\mu}_{i,MK} \geq \hat{\mu}_{i^*,MK}\right)$$
$$= P\left((\hat{\mu}_{i,MK} - \mu_i) \geq (\hat{\mu}_{i^*,MK} - \mu_{i^*}) + \Delta_i\right)$$
$$\leq e^{-\frac{M\Delta_i^2}{4}} \text{ (the difference is } \sqrt{2/M}\text{-sub-Gaussian)}$$

Theorem (Explore-Then-Commit upper bound)

$$\mathbb{E}[\mathcal{R}_T] = \sum_i \Delta_i \mathbb{E}[N_{i,T}] \leq \sum_{i=1}^{K} \Delta_i \left(M + (T - MK)e^{-\frac{M\Delta_i^2}{4}}\right)$$

- For the two arm case, if we know Δ, then $m = \frac{4}{\Delta_1^2} \log \frac{T\Delta_1^2}{4}$, results in $\mathbb{E}[\mathcal{R}_T] \leq \sum_{i=1}^{K} \frac{4}{\Delta_1} \log \frac{T\Delta_1^2}{4} + T\frac{4}{T\Delta_1^2} = O\left(\frac{K\log(T)}{\Delta_1}\right)$
- But we don't know Δ...can we be adaptive?

Algorithm Design Principle: OFU

- OFU: Optimism in the Face of Uncertainty
- We establish some confidence set for the problem instance (e.g. means) to within some confidence set
- We then assume the most favorable instance in the confidence set and act greedily

Algorithm Design Principle: OFU

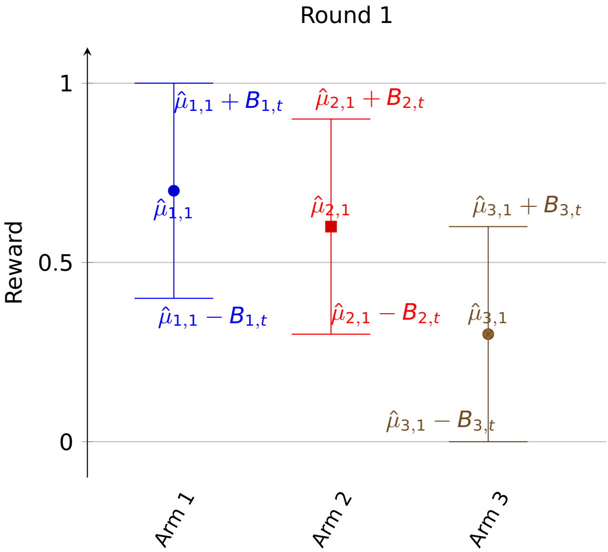
- OFU: Optimism in the Face of Uncertainty
- We establish some confidence set for the problem instance (e.g. means) to within some confidence set
- We then assume the most favorable instance in the confidence set and act greedily

Algorithm: UCB1 [Auer et al., 2002a]

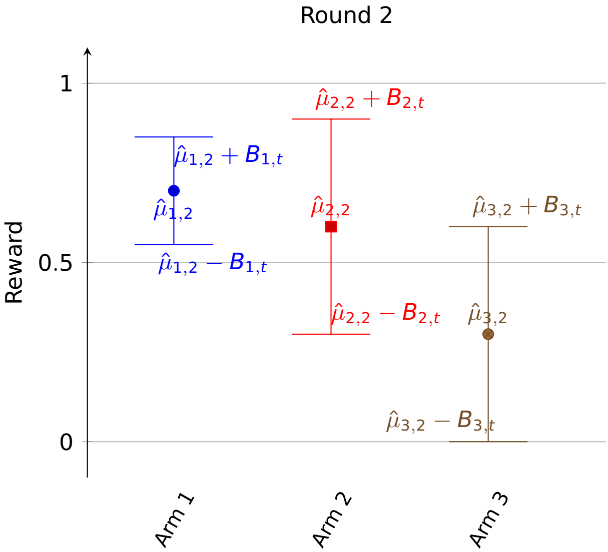
Given: Game length T
Initialize: play every arm once
For $t = K+1, 2, \ldots, T$:

- Compute upper confidence bounds $B_{i,t-1} = \sqrt{\frac{6\log(t)}{N_{i,t-1}}}$
- Choose $I_t = \arg\max_i \hat{\mu}_{i,t-1} + B_{i,t-1}$, observe $X_t \sim \nu_{I_t}$
- Update $N_{i,t} = N_{i,t-1} + \mathbb{1}_{\{I_t=i\}}$ and $\hat{\mu}_{i,t} = \frac{\sum_{s=1}^{t} \mathbb{1}_{\{I_s=i\}} X_s}{N_{i,t}}$

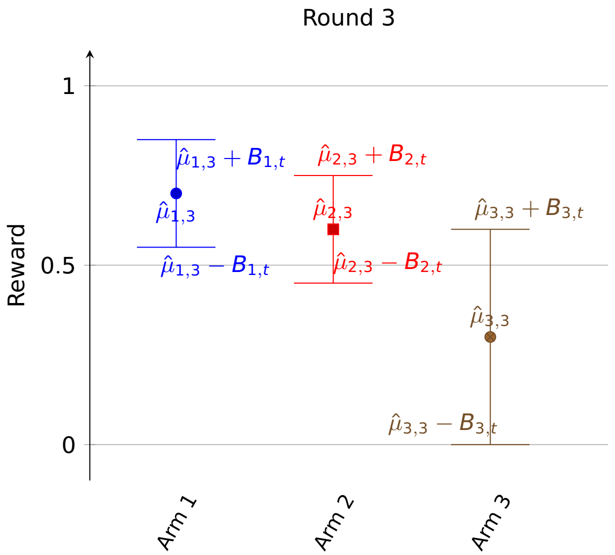
UCB Illustration



UCB Illustration



UCB Illustration



UCB: Intuition

- Naturally balances exploration and exploitation: an arm has a high UCB if
  - It has a high $\hat{\mu}_{i,t}$, or
  - $B_{i,t}$ is large because $N_{i,t-1}$ is small
- Optimistic because we pretend the rewards are the plausibly best and then do the greedy thing

UCB: Analysis

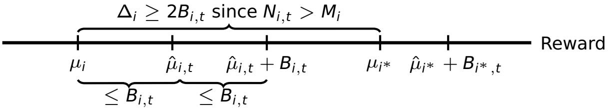
- Define $M_i = \left\lceil \frac{12\log(T)}{\Delta_i^2} \right\rceil$, the number of pulls of arm i such that $B_{i,t} = \sqrt{\frac{6\log(t)}{N_{i,t}}} \leq \sqrt{\frac{6\log(T)}{N_{i,t}}} \leq \frac{\Delta_i}{2}$
- The intuition of the proof is
  1. Since $\overline{\mathcal{R}_T} = \sum_i \Delta_i \mathbb{E}[N_{i,T}]$, we bound $\mathbb{E}[N_{i,t}]$ first.
  2. With high probability, we will never pull arm i more than M_i times, so

$$\mathbb{E}[N_{i,T}] = \mathbb{E}\sum_{t=1}^{T} \mathbb{1}_{\{I_t=i\}} \leq M_i + \sum_{t=M_i}^{T} \underbrace{\mathbb{E}\mathbb{1}_{\{I_t=i,N_{i,t}>M_i\}}}_{\text{we will bound this}}$$

  3. If $\{I_t = i, N_{i,t} > M_i\}$ occurs, then the UCB for i^* or for i must be wrong (next slide)

UCB: Analysis

Claim: if $\{I_t = i, N_{i,t} > M_i\}$ occurs, then either $\hat{\mu}_{i,t}$ must be too high or $\hat{\mu}_{i^*,t}$ must be too low. In a picture:



In an equation: suppose that $N_{i,t} > M_i$, $\hat{\mu}_{i,t} - B_{i,t} < \mu_i$, and $\hat{\mu}_{i^*,t} + B_{i^*,t} > \mu_{i^*}$. Then

$$\hat{\mu}_{i^*,t} + B_{i^*,t} > \mu_{i^*} = \mu_i + \Delta_i \geq \mu_i + \underbrace{2B_{i,t}}_{\text{by choice of } B_{i,t}} > \hat{\mu}_{i,t} + B_{i,t},$$

so the algorithm will not choose $I_t = i$.
If $I_t = i$, at least one of the bounds must be wrong, implying

$$P(I_t = i, N_{i,t} > M_i) \leq P\left(\hat{\mu}_{i,t} \geq \mu_i + B_{i,t}\right) + P\left(\hat{\mu}_{i^*,t} + B_{i^*,t} \leq \mu_{i^*}\right).$$

UCB: Analysis

Using the Hoeffding bound,

$$
\begin{aligned}
P\left(\hat{\mu}_{i,t} - \mu_i \le B_{i,t}\right) &\le P\Bigg(\underbrace{\exists s \le t}_{\text{we don't know } N_{i,t-1}} : \hat{\mu}_{i,s} - \mu_i \le \sqrt{\frac{6\log(t)}{s}}\Bigg) \\
&\le \sum_{s=1}^{t} P\left(\hat{\mu}_{i,s} - \mu_i \le \sqrt{\frac{6\log(t)}{s}}\right) \\
&\le \sum_{s=1}^{t} \exp\left\{-\frac{3\log(t)}{s}\right\} \le \sum_{s=1}^{t} t^{-3} = t^{-2}.
\end{aligned}
$$

UCB: Analysis

Using the Hoeffding bound,

$$
\begin{aligned}
P\left(\hat{\mu}_{i,t} - \mu_i \le B_{i,t}\right) &\le P\Bigg(\underbrace{\exists s \le t}_{\text{we don't know } N_{i,t-1}} : \hat{\mu}_{i,s} - \mu_i \le \sqrt{\frac{6\log(t)}{s}}\Bigg) \\
&\le \sum_{s=1}^{t} P\left(\hat{\mu}_{i,s} - \mu_i \le \sqrt{\frac{6\log(t)}{s}}\right) \\
&\le \sum_{s=1}^{t} \exp\left\{-\frac{3\log(t)}{s}\right\} \le \sum_{s=1}^{t} t^{-3} = t^{-2}.
\end{aligned}
$$

The same inequality holds for i^*, so

$$
\overline{\mathcal{R}_T} = \sum_i \Delta_i \mathbb{E}[N_{i,T}] \le \sum_i \Delta_i \left(\frac{12\log(T)}{\Delta_i^2} + 2\sum_{t=M_i+1}^{T} t^{-2}\right).
$$

UCB: Analysis

Theorem (UCB upper bound [Auer, 2002])

The UCB1 algorithm on 1-sub-Gaussian data has

$$\overline{\mathcal{R}_T} \leq \sum_i \frac{12\log(T)}{\Delta_j} + o(1).$$

UCB: Analysis

Theorem (UCB upper bound [Auer, 2002])

The UCB1 algorithm on 1-sub-Gaussian data has

$$\overline{\mathcal{R}_T} \leq \sum_i \frac{12\log(T)}{\Delta_j} + o(1).$$

Theorem (Lower Bound [Lai and Robbins, 1985])

Suppose we have a parametric family P_θ and $\theta_1, \ldots, \theta_k$. For any "admissible" algorithm,

$$\liminf_{T\to\infty} \frac{\overline{R_T}}{\log(T)} \geq \sum_{i\neq i^*} \frac{\Delta_i}{KL(P_{\theta_i}, P_{\theta_{i^*}})} \approx O\left(\sum_{i\neq i^*} \frac{1}{\Delta_i}\right)$$

E.g. if P_θ is Bernoulli, then $\frac{(\theta_i-\theta_{i^*})^2}{\theta_{i^*}(1-\theta_{i^*})} \geq KL(P_{\theta_i}, P_{\theta_{i^*}}) \geq 2(\theta_i - \theta_{i^*})^2$.

Algorithm Design Principle: Probability Matching

- We put a prior π over means μ_i and a likelihood $\nu_i = P(\cdot|\mu_i)$ over rewards
- Choose $P(I_t = i) = P(\mu_i = \mu_{i^*}|history)$ (the matching)
- We usually pick conjugate models (e.g. $\mu_i \sim N(0,1)$, $X_t \sim N(\mu_i, 1)$)

Algorithm Design Principle: Probability Matching

- We put a prior π over means μ_i and a likelihood $\nu_i = P(\cdot|\mu_i)$ over rewards
- Choose $P(I_t = i) = P(\mu_i = \mu_{i^*}|history)$ (the matching)
- We usually pick conjugate models (e.g. $\mu_i \sim N(0,1)$, $X_t \sim N(\mu_i, 1)$)

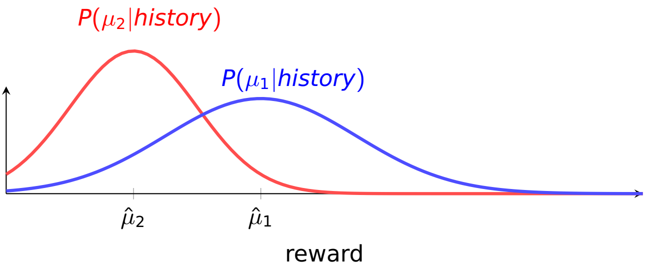
Algorithm: Thompson Sampling

Given: game length T, prior $\pi(\mu)$, likelihoods $p(\cdot|\mu)$
Initialize posteriors $p_{i,0}(\mu) = \pi(\mu)$

For $t = 1, 2, \ldots, T$:

- Draw $\theta_{i,t} \sim p_{i,t-1}$ for all i
- Choose $I_t = \arg\max_i \theta_{i,t}$ (implements the matching)
- Receive and observe $X_t \sim \nu_{I_t}$
- Update the posterior $p_{I_T,t}(\mu) = p(X_t|\mu) p_{I_t,t-1}(\mu)$

Thompson Sampling: Overview



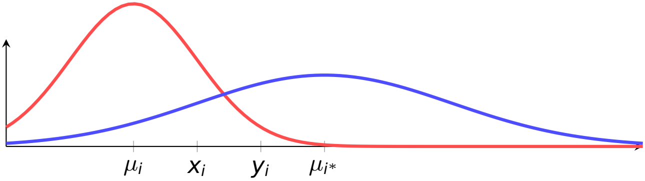
- *Not* Bayesian: a Bayesian method would maximize the Bayes regret (the expectation under the probability model)
- The regret bound is frequentist
- Arms with small $N_{i,t}$ implies a wide posterior, hence a good probability of being selected
- Generally performs empirically better that UCB (it is much more aggressive)
- Analysis is difficult

Thompson Sampling: Upper Bound

Theorem (Agrawal and Goyal [2013])

For binary rewards, Gamma-Beta Thompson sampling has $\mathbb{E}[R_T] \leq (1+\epsilon)\sum_{i \neq i^*} \Delta_i \frac{\log(T)}{KL(\mu_i, \mu_{i^*})} + O\left(\frac{N}{\epsilon^2}\right)$.

- The proof is much more technical that UCB's
- We cannot rely on the upper bounds being correct w.h.p.



For some to-be-tuned $\mu_i \le x_i \le y_i \le \mu_{i^*}$, we have

$$\mathbb{E}[N_{i,T}] \le \sum_{t=1}^{T} P(I_t = i)$$

$$\le \sum_{t=1}^{T} P(I_t = i, \hat{\mu}_{i,t-1} \le x_i, \theta_{i,t} \ge y_i) \qquad (O\left(\frac{\log(T)}{kl(x_j, y_j)}\right))$$

$$+ \sum_{t=1}^{T} P(I_t = i, \hat{\mu}_{i,t-1} \le x_i, \theta_{i,t} \le y_i) \qquad \text{(the tricky case)}$$

$$+ \sum_{t=1}^{T} P(I_t = i, \hat{\mu}_{i,t-1} \ge x_i) \qquad \text{(Small by concentration)}$$

Thompson Sampling: Proof Outline

- The tricky case is $\sum_{t=1}^{T} P(I_t = i, \hat{\mu}_{i,t-1} \le x_i, \theta_{i,t} \le y_i)$
- This happens when we have enough samples of i but not many of i^*
- A key lemma argues that, on $\hat{\mu}_{i,t-1} \le x_i, \theta_{i,t} \le y_i$, the probability of picking i is a constant less than of picking i^*:

$$\sum_{t=1}^{T} P(I_t = i, \hat{\mu}_{i,t-1} \le x_i, \theta_{i,t} \le y_i)$$

$$\le \sum_{t=1}^{T} \underbrace{\frac{P(\theta_{i^*,t} \le y_i)}{P(\theta_{i^*,t} > y_i)}}_{\text{exponentially small}} P(I_t = i^*, \hat{\mu}_{i,t-1} \le x_i, \theta_{i,t} \le y_i) = O(1)$$

- Hence, we will quickly get enough samples of i^*

Best of Both Worlds

- The stochastic and adversarial algorithms are quite different
- A natural question: is there an algorithm that
  - gets $\mathcal{R}_T = O(\sqrt{TK})$ regret for adversarial
  - gets $\mathcal{R}_t = O(\sum_i \log(T)/\Delta_i)$ regret for stochastic
  - without knowing the setting?

Best of Both Worlds

- The stochastic and adversarial algorithms are quite different
- A natural question: is there an algorithm that
  - gets $\mathcal{R}_T = O(\sqrt{TK})$ regret for adversarial
  - gets $\mathcal{R}_t = O(\sum_i \log(T)/\Delta_i)$ regret for stochastic
  - without knowing the setting?
- Bubeck and Slivkins [2012] proposed an algorithm that assumes stochastic but falls back to UCB once adversarial data is detected
- Zimmert and Seldin [2019] showed that (for pseudo-regret), it is possible
  - Their algorithm: online mirror descent with $\frac{1}{2}$-Tsallis entropy
  - $\Psi(w) = -\sum_i 4(\sqrt{w_i} - \frac{1}{2} w_i)$

A New Problem

- What if we only wanted to identify the best arm i^* without caring about loss along the way?
- Intuitively, we would explore more; we are happy to accrue less reward if we get more useful samples.
- More similar to hypothesis testing; useful for selecting treatments
- Known as "Best Arm Identification" or "Pure Exploration"

Two Settings

Protocol: Best-arm Identification

Given:number of arms K, arm distributions $\nu_1, \ldots, \nu_K$

For $t = 1, 2, \ldots,$

- The learner picks arm $I_t \in \{1, \ldots, K\}$
- The learner observes $X_t \sim \nu_{I_t}$
- The learner decides whether to stop

The learner returns arm A

Two settings:

	fixed-confidence	*fixed-budget*
Input	$\delta > 0$,	T
Goal	$P(A = i^*) \geq 1 - \delta$	maximize $P(A = i^*)$
Stopping	once learner is confident	after T rounds

- Standard stochastic bandit algorithms under explore (they fail to meet lower bounds on this problem)
- Many can be adapted
  - LUCB [Kalyanakrishnan et al., 2012]
  - Top-Two Thompson Samping [Russo, 2016]
- Instead, we will describe a new algorithm design principle

Algorithm Design Principle: Action Elimination

Algorithm: Successive Elimination

Given: confidence $\delta > 0$
Initialize plausibly-best set $S = \{1, \ldots, K\}$

For $t = 1, 2, \ldots$:

- Pull all arms in S and update $\hat{\mu}_{i,t}$
- Calculate $B_t = \sqrt{2t^{-1} \log(4Kt^2/\delta)}$
- Remove i from S if $\underbrace{\max_{j \in S} \hat{\mu}_{j,t} - B_t}_{\text{Lowest } \mu_i^* \text{ could be}} \geq \underbrace{\hat{\mu}_{i,t} + B_t}_{\text{highest } \mu_i \text{ could be}}$
- If $|S| = 1$, stop and return $A = S$.

- S is a list of plausibly-best arms
- Each epoch, all arms that cannot be the best (if the bounds hold) are removed

Successive Elimination Analysis

- Define the "bad event" $\mathcal{E} = \bigcup_{i,t}\{|\hat{\mu}_{i,t} - \mu_i| \geq B_t(\delta)\}$: we have

$$P(\mathcal{E}) \leq \sum_{i,t} P\left(|\hat{\mu}_{i,t} - \mu_i| \geq \sqrt{2t^{-1}\log(4Kt^2/\delta)}\right) \leq \sum_{i,t} 2e^{-\log\left(\frac{4Kt^2}{\delta}\right)}$$

$$\leq \sum_{i,t} \frac{2\delta}{4Kt^2} = \frac{2\pi^2}{24}\delta \leq \delta$$

Successive Elimination Analysis

- Define the "bad event" $\mathcal{E} = \bigcup_{i,t}\{|\hat{\mu}_{i,t} - \mu_i| \geq B_t(\delta)\}$: we have

$$P(\mathcal{E}) \leq \sum_{i,t} P\left(|\hat{\mu}_{i,t} - \mu_i| \geq \sqrt{2t^{-1}\log(4Kt^2/\delta)}\right) \leq \sum_{i,t} 2e^{-\log\left(\frac{4Kt^2}{\delta}\right)}$$

$$\leq \sum_{i,t} \frac{2\delta}{4Kt^2} = \frac{2\pi^2}{24}\delta \leq \delta$$

- (Correctness) If $\mathcal{E}$ does not happen,
  - $|\hat{\mu}_{i^*} - \mu_{i^*}| \leq B_t$ and $|\mu_j - \hat{\mu}_j| \leq B_t$ for all j. Thus, for all j $\hat{\mu}_j - \hat{\mu}_{i^*} \leq (\mu_{i^*} - \hat{\mu}_{i^*}) + (\mu_j - \mu_{i^*}) + (\hat{\mu}_j - \mu_j) \leq 2B_t$
  - i is removed if $\max_{j \in S} \hat{\mu}_{j,t} - \hat{\mu}_{i,t} \geq 2B_t \Rightarrow i^*$ is never removed
  - $\lim_{t\to\infty} B_t(\delta) \to 0$: every arm will eventually be removed
  - Successive Elimination is correct with probability $1 - \delta$

Successive Elimination Analysis

- Define the "bad event" $\mathcal{E} = \bigcup_{i,t}\{|\hat{\mu}_{i,t} - \mu_i| \geq B_t(\delta)\}$: we have

$$P(\mathcal{E}) \leq \sum_{i,t} P\left(|\hat{\mu}_{i,t} - \mu_i| \geq \sqrt{2t^{-1}\log(4Kt^2/\delta)}\right) \leq \sum_{i,t} 2e^{-\log\left(\frac{4Kt^2}{\delta}\right)}$$

$$\leq \sum_{i,t} \frac{2\delta}{4Kt^2} = \frac{2\pi^2}{24}\delta \leq \delta$$

- (Correctness) If $\mathcal{E}$ does not happen,
  - $|\hat{\mu}_{i^*} - \mu_{i^*}| \leq B_t$ and $|\mu_j - \hat{\mu}_j| \leq B_t$ for all j. Thus, for all j $\hat{\mu}_j - \hat{\mu}_{i^*} \leq (\mu_{i^*} - \hat{\mu}_{i^*}) + (\mu_j - \mu_{i^*}) + (\hat{\mu}_j - \mu_j) \leq 2B_t$
  - i is removed if $\max_{j \in S} \hat{\mu}_{j,t} - \hat{\mu}_{i,t} \geq 2B_t \Rightarrow i^*$ is never removed
  - $\lim_{t\to\infty} B_t(\delta) \to 0$: every arm will eventually be removed
  - Successive Elimination is correct with probability $1 - \delta$
- (Sample Complexity): arm i will be eliminated once $\Delta_i \leq 2B_t$
  - We can verify that $N_i = O\left(\Delta_i^{-2}\log(K/\delta\Delta_i)\right)$ is sufficient
  - Total sample complexity of $\sum_i \Delta_i^{-2}\log(K/\delta\Delta_i)$

Theorem

Successive Elimination is $(0, \delta)$-PAC with sample complexity

$$O\left(\sum_i \Delta_i^{-2} \log(K/\delta\Delta_i)\right)$$

Theorem

For any best-arm identification algorithm, there is a problem instance that requires

$$\Omega\left(\sum_i \Delta_i^{-2} \log\log\left(\frac{1}{\delta\Delta_i^2}\right)\right)$$

samples.

Bonus: Linear Contextual Bandits

Protocol: Contextual Linear Bandit

Given: game length T, number of arms K

For $t = 1, 2, \ldots, T$,

- The learner sees one context per arm $c_{1,t}, \ldots, c_{K,t}$
- The learner picks action $I_t \in \{1, \ldots, K\}$
- The learner observes and receives reward $X_t = \langle c_{I_t,t}, \theta^* \rangle + \xi_t$

Regret is defined w.r.t. an agent that knows the true θ:

$$\overline{\mathcal{R}}_T = \sum_{t=1}^{T} \max_i c_{i,t}^\intercal \theta^* - \sum_{t=1}^{T} c_{I_t,t}^\intercal \theta^*$$

Algorithm Design Principle: Optimism

Algorithm: OFUL [Abbasi-Yadkori et al., 2011]

Initialize $\hat{\theta}_0 = 0$, $B_0 = \mathbb{R}^d$
For $t = 1, 2, \ldots, T$:

- Receive contexts $c_{1,t}, \ldots, c_{K,t}$
- Choose $(I_t, \tilde{\theta}_t) = \arg\max_{i \in \{1,\ldots,K\}, \theta \in B_{t-1}} \theta^\intercal c_{i,t}$ (optimism)
- Observe $X_t = c_{I_t,t}^\intercal \theta^* + \xi_t$
- Calculate $V_t = \sum_{s=1}^t c_s c_s^\intercal + \lambda I$ and $r_t = \sqrt{\log \frac{\det(V_t)}{\delta^2 \lambda^d}} + \sqrt{\lambda}\|\theta^*\|$
- Calculate $\hat{\theta}_t = V_t^{-1} \left(\sum_{s=1}^t c_s X_s\right)$ (ridge)
- Update $B_t = \{\theta : (\theta - \hat{\theta}_t)^\intercal V_t (\theta - \hat{\theta}_t) \le r_t)$

- If ξ_t is 1-sub-Gaussian, B_t is a confidence sequence with $P(\forall t > 0 : \theta^* \in \boldsymbol{B}_t) \ge 1 - \delta$ (more examples in [de la Peña et al., 2009, Howard et al., 2020])

Analysis

- Regret decomposes over rounds:
- Recall that $(I_t, \tilde{\theta}_t) = \arg\max_{i \in \{1,\ldots,K\}, \theta \in B_{t-1}} \theta^\intercal c_{i,t}$

$$
\begin{aligned}
\mathcal{R}_t - \mathcal{R}_{t-1} &= c_{i_t^*}^\intercal \theta^* - c_{I_t}^\intercal \theta^* \\
&\leq c_{I_t}^\intercal \tilde{\theta}_t - c_{I_t}^\intercal \theta^* && \text{(by optimism)} \\
&\leq c_{I_t}^\intercal \left(\tilde{\theta}_t - \hat{\theta}_{t-1} \right) + c_{I_t}^\intercal \left(\hat{\theta}_{t-1} - \theta^* \right) \\
&\leq \|c_{I_t}\|_{V_t} \underbrace{\left\| \tilde{\theta}_t - \hat{\theta}_{t-1} \right\|_{V_t}}_{\leq r_t} + \|c_{I_t}\|_{V_t} \underbrace{\left\| \hat{\theta}_{t-1} - \theta^* \right\|_{V_t}}_{\leq r_t}
\end{aligned}
$$

- After some algebra, we can show, with probability $\geq 1 - \delta$, that

$$
\mathcal{R}_T = O\left(\frac{d \log(1/\delta)}{\Delta} \right)
$$

- The shared structure lets us learn a lot!

Review

- Setting: adversarial bandits
  - Exp3 (exponential weights)
- Setting: stochastic bandits
  - UCB (optimism)
  - Thompson Sampling (probablity matching)
- Setting: pure exploration
  - Successive Elimination (action-elimination)
- Setting: linear contextual bandits
  - OFUL (optimism)

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *Artificial intelligence and statistics*, pages 99–107, 2013.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. 2009.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002a.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, (1): 48–77, 2002b.

Gábor Bartók, Dean P Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring – classification, regret bounds, and algorithms. *Mathematics of Operations Research*, (4):967–997, 2014.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42–1, 2012.

Victor H. de la Peña, Michael J. Klass, and Tze Leung Lai. Theory and applications of multivariate self-normalized processes. *Stochastic Processes and their Applications*, 119(12):4210–4227, December 2009. ISSN 0304-4149.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probab. Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321. URL https://doi.org/10.1214/18-PS321.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662, 2012.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, (1):4–22, 1985.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems*, pages 684–692, 2011.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3168–3176, 2015.

Daniel Russo. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418, 2016.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 467–475, 2019.

Aside: Lower Bound Reasoning

- Fix a strategy and consider two problem instances:
  1. $\nu_1, \nu_2, \ldots, \nu_K$; with P as the joint distribution over $(I_t, r_{i,t})$
  2. $\nu_1, \nu_2', \ldots, \nu_K$; with P' as the joint distribution over $(I_t, r_{i,t})$
  3. The optimal arm is different: $\mu_2' \geq \mu_1 \geq \mu_2 \geq \mu_3 \geq \ldots$
  4. The data from P and P' will look very similar
- An algorithm that does well on P must not pull arm 2 too many times; hence, it will not do well on P'
- "Similar" is quantified by a change-of-measure identity; e.g. $P'(A) = e^{-\widehat{kl}_{N_{2,T}}} P(A)$, where $\widehat{kl}_t = \sum_{s=1}^{t} \log \frac{d\nu_2}{d\nu_2'}(X_{2,s})$
- Hence, an algorithm cannot tell if it is P or P' and must get high regret under P', mistakenly believing it is playing in P