CSE 541: Interactive Learning

Lecture 8: EXP3, and why $\sqrt{nT}$ and $\log T/\Delta$ are the right answers

Kevin Jamieson

Paul G. Allen School of
Computer Science & Engineering
University of Washington

Monday, October 26, 2026

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Last time: three stochastic bandit algorithms

n arms, 1-sub-Gaussian rewards with means θ_i^*, gaps Δ_i, T_i pulls of arm i, regret $R_T = \sum_i \Delta_i \mathbb{E}[T_i]$.

- **Successive elimination:** pull survivors equally; drop arms confidently worse than the leader; precision doubles each round.
- **UCB:** play $\arg\max_i \ \hat{\theta}_i(t) + \sqrt{2\log(2nT^2/\delta)/T_i(t)}$. Optimism: if a bad arm is played, its interval is still wide, so $T_i \lesssim \Delta_i^{-2}\log(nT)$.
- **Thompson sampling:** sample $\tilde{\theta}_i \sim \mathcal{N}(\hat{\theta}_i, 1/T_i)$ for each arm; play $\arg\max_i \tilde{\theta}_i$. Probability matching; same rates.

All three

$$R_T \ \lesssim \ \sum_{i \geq 2} \frac{\log T}{\Delta_i} \qquad \text{and} \qquad R_T \ \lesssim \ \sqrt{nT \log T},$$

the second via $\min\{\Delta_i T_i, \ \log T/\Delta_i\} \leq \sqrt{T_i \log T}$ and Cauchy–Schwarz.

Last time: three stochastic bandit algorithms

n arms, 1-sub-Gaussian rewards with means θ_i^*, gaps Δ_i, T_i pulls of arm i, regret $R_T = \sum_i \Delta_i \mathbb{E}[T_i]$.

- **Successive elimination:** pull survivors equally; drop arms confidently worse than the leader; precision doubles each round.
- **UCB:** play $\arg\max_i \ \hat{\theta}_i(t) + \sqrt{2\log(2nT^2/\delta)/T_i(t)}$. Optimism: if a bad arm is played, its interval is still wide, so $T_i \lesssim \Delta_i^{-2}\log(nT)$.
- **Thompson sampling:** sample $\tilde{\theta}_i \sim \mathcal{N}(\hat{\theta}_i, 1/T_i)$ for each arm; play $\arg\max_i \tilde{\theta}_i$. Probability matching; same rates.

All three

$$R_T \ \lesssim \ \sum_{i \geq 2} \frac{\log T}{\Delta_i} \qquad \text{and} \qquad R_T \ \lesssim \ \sqrt{nT \log T},$$

the second via $\min\{\Delta_i T_i, \ \log T/\Delta_i\} \leq \sqrt{T_i \log T}$ and Cauchy–Schwarz.

Today. (1) Drop i.i.d.: an adversary picks the losses, and EXP3 still gets $\sqrt{nT\log n}$. (2) Why nobody will beat $\sqrt{nT}$ or $\sum_i 2\log T/\Delta_i$: two constructions, main ideas only.

Adversarial bandits

Exponential weights, with one number per round

Adversarial bandits: the setting

Adversarial multi-armed bandit (notes, Ch. 7)

adversary fixes losses $\ell_1, \ldots, \ell_T \in [-1, 1]^n$ in advance (*oblivious*)
for $t = 1, 2, \ldots, T$:
    player picks $q_t \in \triangle_n$ and plays $I_t \sim q_t$
    player suffers and observes ℓ_{t,I_t}, and nothing else

$\mathrm{Regret}_T = \max_{i \in [n]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,I_t} - \ell_{t,i}\right]$, the expectation over the player's coins only.

Where this sits

Experts (Lecture 2): adversarial losses, but you see all of ℓ_t. EW gets $\sqrt{T \log n / 2}$.

Stochastic bandits (Lectures 5–7): one number per round, but i.i.d. noise around fixed means.

Now both: one number per round, chosen by an adversary.

Adversarial bandits: the setting

Adversarial multi-armed bandit (notes, Ch. 7)

adversary fixes losses $\ell_1, \ldots, \ell_T \in [-1, 1]^n$ in advance (*oblivious*)
for $t = 1, 2, \ldots, T$:
  player picks $q_t \in \triangle_n$ and plays $I_t \sim q_t$
  player suffers and observes ℓ_{t,I_t}, and nothing else

$\mathrm{Regret}_T = \max_{i \in [n]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,I_t} - \ell_{t,i}\right]$, the expectation over the player's coins only.

Where this sits

Experts (Lecture 2): adversarial losses, but you see all of ℓ_t. EW gets $\sqrt{T \log n / 2}$.

Stochastic bandits (Lectures 5–7): one number per round, but i.i.d. noise around fixed means.

Now both: one number per round, chosen by an adversary.

- Randomization is essential, as in Lecture 2: a deterministic player can be made to pay 1 every round while some arm pays ≤ 0 on average.
- Exponential weights is the natural algorithm, but its update $w_{t+1,i} = w_{t,i} e^{-\eta \ell_{t,i}}$ needs $\ell_{t,i}$ for *every* i; we see one coordinate.

The trick: estimate the whole loss vector from one coordinate

Play $I_t \sim q_t$ and define the importance-weighted estimate

$$\hat{\ell}_{t,i} \;=\; \frac{\mathbf{1}\{I_t = i\}}{q_{t,i}}\,\ell_{t,i} \qquad (i \in [n]),$$

which is computable: it is 0 for every arm except I_t, and $\ell_{t,I_t}/q_{t,I_t}$ there.

The trick: estimate the whole loss vector from one coordinate

Play $I_t \sim q_t$ and define the importance-weighted estimate

$$\hat{\ell}_{t,i} \;=\; \frac{\mathbf{1}\{I_t = i\}}{q_{t,i}}\,\ell_{t,i} \qquad (i \in [n]),$$

which is computable: it is 0 for every arm except I_t, and $\ell_{t,I_t}/q_{t,I_t}$ there.

Unbiased. Conditionally on q_t, $\mathbb{E}[\hat{\ell}_{t,i} \mid q_t] =$

The trick: estimate the whole loss vector from one coordinate

Play $I_t \sim q_t$ and define the importance-weighted estimate

$$\hat{\ell}_{t,i} \;=\; \frac{\mathbf{1}\{I_t = i\}}{q_{t,i}}\,\ell_{t,i} \qquad (i \in [n]),$$

which is computable: it is 0 for every arm except I_t, and $\ell_{t,I_t}/q_{t,I_t}$ there.

Unbiased. Conditionally on q_t, $\mathbb{E}[\hat{\ell}_{t,i} \mid q_t] = \sum_{j=1}^{n} q_{t,j}\,\frac{\mathbf{1}\{j=i\}}{q_{t,i}}\,\ell_{t,i} = \ell_{t,i}$.

The trick: estimate the whole loss vector from one coordinate

Play $I_t \sim q_t$ and define the importance-weighted estimate

$$\hat{\ell}_{t,i} \;=\; \frac{\mathbf{1}\{I_t = i\}}{q_{t,i}}\,\ell_{t,i} \qquad (i \in [n]),$$

which is computable: it is 0 for every arm except I_t, and $\ell_{t,I_t}/q_{t,I_t}$ there.

Unbiased. Conditionally on q_t, $\mathbb{E}[\hat{\ell}_{t,i} \mid q_t] = \sum_{j=1}^{n} q_{t,j}\,\frac{\mathbf{1}\{j=i\}}{q_{t,i}}\,\ell_{t,i} = \ell_{t,i}$.

But noisy. Its second moment is $\mathbb{E}[\hat{\ell}_{t,i}^2 \mid q_t] =$
which blows up for arms we rarely play. So we will never let $q_{t,i}$ get too small: mix in a little exploration.

The trick: estimate the whole loss vector from one coordinate

Play $I_t \sim q_t$ and define the importance-weighted estimate

$$\hat{\ell}_{t,i} \;=\; \frac{\mathbf{1}\{I_t = i\}}{q_{t,i}}\,\ell_{t,i} \qquad (i \in [n]),$$

which is computable: it is 0 for every arm except I_t, and $\ell_{t,I_t}/q_{t,I_t}$ there.

Unbiased. Conditionally on q_t, $\mathbb{E}[\hat{\ell}_{t,i} \mid q_t] = \sum_{j=1}^{n} q_{t,j}\,\frac{\mathbf{1}\{j=i\}}{q_{t,i}}\,\ell_{t,i} = \ell_{t,i}$.

But noisy. Its second moment is $\mathbb{E}[\hat{\ell}_{t,i}^2 \mid q_t] = \sum_{j=1}^{n} q_{t,j}\,\frac{\mathbf{1}\{j=i\}}{q_{t,i}^2}\,\ell_{t,i}^2 = \frac{\ell_{t,i}^2}{q_{t,i}}$, which blows up for arms we rarely play. So we will never let $q_{t,i}$ get too small: mix in a little exploration.

The trick: estimate the whole loss vector from one coordinate

Play $I_t \sim q_t$ and define the importance-weighted estimate

$$\hat{\ell}_{t,i} \;=\; \frac{\mathbf{1}\{I_t = i\}}{q_{t,i}}\,\ell_{t,i} \qquad (i \in [n]),$$

which is computable: it is 0 for every arm except I_t, and $\ell_{t,I_t}/q_{t,I_t}$ there.

Unbiased. Conditionally on q_t, $\mathbb{E}[\hat{\ell}_{t,i} \mid q_t] = \sum_{j=1}^{n} q_{t,j}\,\frac{\mathbf{1}\{j=i\}}{q_{t,i}}\,\ell_{t,i} = \ell_{t,i}$.

But noisy. Its second moment is $\mathbb{E}[\hat{\ell}_{t,i}^2 \mid q_t] = \sum_{j=1}^{n} q_{t,j}\,\frac{\mathbf{1}\{j=i\}}{q_{t,i}^2}\,\ell_{t,i}^2 = \frac{\ell_{t,i}^2}{q_{t,i}}$, which blows up for arms we rarely play. So we will never let $q_{t,i}$ get too small: mix in a little exploration.

EXP3(γ): exponential weights for exploration and exploitation

Input: $\eta > 0$, $\gamma \in [0,1]$, exploration distribution $\lambda \in \triangle_n$ (e.g. uniform). **Initialize** $w_1 \equiv 1$.

for $t = 1, \ldots, T$: $p_t = w_t / \sum_j w_{t,j}$; play $I_t \sim q_t := (1-\gamma)p_t + \gamma\lambda$; observe ℓ_{t,I_t};

form $\hat{\ell}_t$ as above; $w_{t+1,i} = w_{t,i}\, e^{-\eta \hat{\ell}_{t,i}}$ for all i.

Exponential weights, run on the *estimated* losses, with its play smoothed toward λ.

The regret of EXP3: a general theorem

Theorem (notes, Thm. 12)

Fix any losses $\ell_t \in [-1,1]^n$. For any estimates with $\mathbb{E}[\hat{\ell}_{t,i} \mid \mathcal{F}_{t-1}] = \ell_{t,i}$ and $-\eta\hat{\ell}_{t,i} \le 1$ for all i, t,

$$\max_i \mathbb{E}\Big[\sum_{t=1}^{T} \ell_{t,I_t} - \ell_{t,i}\Big] \;\le\; 2\gamma T + \frac{\log n}{\eta} + (1-\gamma)\,\eta\,\mathbb{E}\Big[\sum_{t=1}^{T}\sum_{j=1}^{n} p_{t,j}\hat{\ell}_{t,j}^2\Big].$$

The regret of EXP3: a general theorem

Theorem (notes, Thm. 12)

Fix any losses $\ell_t \in [-1,1]^n$. For any estimates with $\mathbb{E}[\hat{\ell}_{t,i} \mid \mathcal{F}_{t-1}] = \ell_{t,i}$ and $-\eta\hat{\ell}_{t,i} \le 1$ for all i, t,

$$\max_i \mathbb{E}\Big[\sum_{t=1}^{T} \ell_{t,I_t} - \ell_{t,i}\Big] \;\le\; 2\gamma T + \frac{\log n}{\eta} + (1-\gamma)\,\eta\,\mathbb{E}\Big[\sum_{t=1}^{T}\sum_{j=1}^{n} p_{t,j}\hat{\ell}_{t,j}^2\Big].$$

Step A: pass to the estimated losses. Fix i. Since $q_t = (1-\gamma)p_t + \gamma\lambda$ and $|\ell_{t,j}| \le 1$,

$$\mathbb{E}\Big[\sum_t \ell_{t,I_t} - \ell_{t,i}\Big] = \mathbb{E}\Big[\sum_t \langle q_t, \ell_t\rangle - \ell_{t,i}\Big] \;\le$$

$$\le$$

The regret of EXP3: a general theorem

Theorem (notes, Thm. 12)

Fix any losses $\ell_t \in [-1,1]^n$. For any estimates with $\mathbb{E}[\hat{\ell}_{t,i} \mid \mathcal{F}_{t-1}] = \ell_{t,i}$ and $-\eta\hat{\ell}_{t,i} \le 1$ for all i, t,

$$\max_i \mathbb{E}\Big[\sum_{t=1}^{T} \ell_{t,I_t} - \ell_{t,i}\Big] \;\le\; 2\gamma T + \frac{\log n}{\eta} + (1-\gamma)\,\eta\,\mathbb{E}\Big[\sum_{t=1}^{T}\sum_{j=1}^{n} p_{t,j}\hat{\ell}_{t,j}^2\Big].$$

Step A: pass to the estimated losses. Fix i. Since $q_t = (1-\gamma)p_t + \gamma\lambda$ and $|\ell_{t,j}| \le 1$,

$$\mathbb{E}\Big[\sum_t \ell_{t,I_t} - \ell_{t,i}\Big] = \mathbb{E}\Big[\sum_t \langle q_t, \ell_t\rangle - \ell_{t,i}\Big] \;\le\; \gamma T + \mathbb{E}\Big[\sum_t (1-\gamma)\langle p_t, \ell_t\rangle - \ell_{t,i}\Big]$$
$$\le\; 2\gamma T + (1-\gamma)\,\mathbb{E}\Big[\sum_t \langle p_t, \ell_t\rangle - \ell_{t,i}\Big],$$

using $\gamma\langle \lambda, \ell_t\rangle \le \gamma$ and $-\ell_{t,i} \le -(1-\gamma)\ell_{t,i} + \gamma$. As p_t depends only on the past and $\mathbb{E}[\hat{\ell}_t \mid \mathcal{F}_{t-1}] = \ell_t$, the last term is $\mathbb{E}\big[\sum_t \langle p_t, \hat{\ell}_t\rangle - \hat{\ell}_{t,i}\big]$: EW's regret *on the estimates*.

Step B: the potential argument

$W_t = \sum_i w_{t,i}$, $W_1 = n$. **Lower bound**, as always:
$\log \frac{W_{T+1}}{W_1} \geq \log w_{T+1,i} - \log n = -\eta \sum_t \hat{\ell}_{t,i} - \log n.$

Step B: the potential argument

$W_t = \sum_i w_{t,i}$, $W_1 = n$. **Lower bound**, as always:
$\log \frac{W_{T+1}}{W_1} \geq \log w_{T+1,i} - \log n = -\eta \sum_t \hat{\ell}_{t,i} - \log n$.

Upper bound. Hoeffding's lemma (slack $\eta^2/8$ per round) needed bounded losses; $\hat{\ell}_{t,i}$ can be huge. What we have is $-\eta\hat{\ell}_{t,i} \leq 1$, and $e^x \leq 1 + x + x^2$ for $x \leq 1$. One round:

$$\log \frac{W_{t+1}}{W_t} = \log \left(\sum_i p_{t,i}\, e^{-\eta \hat{\ell}_{t,i}} \right) \leq$$

$$\leq$$

Step B: the potential argument

$W_t = \sum_i w_{t,i}$, $W_1 = n$. **Lower bound**, as always:
$\log \frac{W_{T+1}}{W_1} \geq \log w_{T+1,i} - \log n = -\eta \sum_t \hat{\ell}_{t,i} - \log n$.

Upper bound. Hoeffding's lemma (slack $\eta^2/8$ per round) needed bounded losses; $\hat{\ell}_{t,i}$ can be huge. What we have is $-\eta\hat{\ell}_{t,i} \leq 1$, and $e^x \leq 1 + x + x^2$ for $x \leq 1$. One round:

$$\log \frac{W_{t+1}}{W_t} = \log \Big(\sum_i p_{t,i}\, e^{-\eta \hat{\ell}_{t,i}} \Big) \leq \log \Big(1 - \eta \langle p_t, \hat{\ell}_t \rangle + \eta^2 \sum_i p_{t,i} \hat{\ell}_{t,i}^2 \Big)$$
$$\leq -\eta \langle p_t, \hat{\ell}_t \rangle + \eta^2 \sum_i p_{t,i} \hat{\ell}_{t,i}^2$$

by $e^x \leq 1 + x + x^2$, then $\log(1+x) \leq x$. Sum over t and combine with the lower bound:

$$\sum_{t=1}^{T} \langle p_t, \hat{\ell}_t \rangle - \sum_{t=1}^{T} \hat{\ell}_{t,i} \;\leq\; \frac{\log n}{\eta} + \eta \sum_{t=1}^{T} \sum_{j=1}^{n} p_{t,j} \hat{\ell}_{t,j}^2 .$$

Take expectations and plug into Step A. The new slack, $\eta^2 \sum_j p_{t,j} \hat{\ell}_{t,j}^2$, is what exploration keeps in check. ■

EXP3 for n-armed bandits: $\sqrt{12\,nT\log n}$

Corollary (notes, Cor. 3)

With λ uniform, $\gamma = \eta n$ and $\eta = \sqrt{\log(n)/(3nT)}$:

$$\max_i \mathbb{E}\Big[\sum_{t=1}^{T} \ell_{t,I_t} - \ell_{t,i}\Big] \;\le\; \sqrt{12\,nT\log n}.$$

EXP3 for n-armed bandits: $\sqrt{12\,nT\log n}$

Corollary (notes, Cor. 3)

With λ uniform, $\gamma = \eta n$ and $\eta = \sqrt{\log(n)/(3nT)}$:

$$\max_i \mathbb{E}\Big[\sum_{t=1}^{T} \ell_{t,I_t} - \ell_{t,i}\Big] \;\le\; \sqrt{12\,nT\log n}.$$

The condition. Since $q_{t,i} \ge \gamma\lambda_i = \gamma/n$ and $\gamma = \eta n$,
$|\eta\hat{\ell}_{t,i}| = \eta\,\frac{\mathbf{1}\{I_t=i\}}{q_{t,i}}\,|\ell_{t,i}| \le$

EXP3 for n-armed bandits: $\sqrt{12\,nT\log n}$

Corollary (notes, Cor. 3)

With λ uniform, $\gamma = \eta n$ and $\eta = \sqrt{\log(n)/(3nT)}$:

$$\max_i \mathbb{E}\Big[\sum_{t=1}^{T} \ell_{t,I_t} - \ell_{t,i}\Big] \;\le\; \sqrt{12\,nT\log n}.$$

The condition. Since $q_{t,i} \ge \gamma\lambda_i = \gamma/n$ and $\gamma = \eta n$,
$|\eta\hat{\ell}_{t,i}| = \eta\,\frac{\mathbf{1}\{I_t=i\}}{q_{t,i}}\,|\ell_{t,i}| \le \frac{\eta n}{\gamma} = 1.$

EXP3 for n-armed bandits: $\sqrt{12\,nT\log n}$

Corollary (notes, Cor. 3)

With λ uniform, $\gamma = \eta n$ and $\eta = \sqrt{\log(n)/(3nT)}$:

$$\max_i \mathbb{E}\Big[\sum_{t=1}^{T} \ell_{t,I_t} - \ell_{t,i}\Big] \;\le\; \sqrt{12\,nT\log n}.$$

The condition. Since $q_{t,i} \ge \gamma\lambda_i = \gamma/n$ and $\gamma = \eta n$,
$|\eta\hat{\ell}_{t,i}| = \eta\,\frac{\mathbf{1}\{I_t=i\}}{q_{t,i}}\,|\ell_{t,i}| \le \frac{\eta n}{\gamma} = 1.$

The second moment. Only $j = I_t$ contributes, and $p_{t,j} \le q_{t,j}/(1-\gamma)$:

$$\mathbb{E}\Big[\sum_{j=1}^{n} p_{t,j}\hat{\ell}_{t,j}^2\Big] = \mathbb{E}\Big[p_{t,I_t}\frac{\ell_{t,I_t}^2}{q_{t,I_t}^2}\Big] \;\le$$

EXP3 for n-armed bandits: $\sqrt{12\,nT\log n}$

Corollary (notes, Cor. 3)

With λ uniform, $\gamma = \eta n$ and $\eta = \sqrt{\log(n)/(3nT)}$:

$$\max_i \mathbb{E}\Big[\sum_{t=1}^{T} \ell_{t,I_t} - \ell_{t,i}\Big] \;\le\; \sqrt{12\,nT\log n}.$$

The condition. Since $q_{t,i} \ge \gamma\lambda_i = \gamma/n$ and $\gamma = \eta n$,
$|\eta\hat{\ell}_{t,i}| = \eta\,\frac{\mathbf{1}\{I_t=i\}}{q_{t,i}}\,|\ell_{t,i}| \le \frac{\eta n}{\gamma} = 1.$

The second moment. Only $j = I_t$ contributes, and $p_{t,j} \le q_{t,j}/(1-\gamma)$:

$$\mathbb{E}\Big[\sum_{j=1}^{n} p_{t,j}\hat{\ell}_{t,j}^2\Big] = \mathbb{E}\Big[p_{t,I_t}\frac{\ell_{t,I_t}^2}{q_{t,I_t}^2}\Big] \;\le\; \frac{1}{1-\gamma}\,\mathbb{E}\Big[\frac{\ell_{t,I_t}^2}{q_{t,I_t}}\Big] = \frac{1}{1-\gamma}\sum_{j=1}^{n}\ell_{t,j}^2 \;\le\; \frac{n}{1-\gamma}.$$

So $\mathrm{Regret}_T \le 3\eta nT + \log(n)/\eta = \sqrt{12\,nT\log n}$ at $\eta = \sqrt{\log(n)/(3nT)}$. ■

EXP3 for n-armed bandits: $\sqrt{12\,nT\log n}$

Corollary (notes, Cor. 3)

With λ uniform, $\gamma = \eta n$ and $\eta = \sqrt{\log(n)/(3nT)}$:

$$\max_i \mathbb{E}\Big[\sum_{t=1}^{T} \ell_{t,I_t} - \ell_{t,i}\Big] \;\le\; \sqrt{12\,nT\log n}.$$

The condition. Since $q_{t,i} \ge \gamma\lambda_i = \gamma/n$ and $\gamma = \eta n$,
$|\eta\hat{\ell}_{t,i}| = \eta\,\frac{\mathbf{1}\{I_t=i\}}{q_{t,i}}\,|\ell_{t,i}| \le \frac{\eta n}{\gamma} = 1.$

The second moment. Only $j = I_t$ contributes, and $p_{t,j} \le q_{t,j}/(1-\gamma)$:

$$\mathbb{E}\Big[\sum_{j=1}^{n} p_{t,j}\hat{\ell}_{t,j}^2\Big] = \mathbb{E}\Big[p_{t,I_t}\frac{\ell_{t,I_t}^2}{q_{t,I_t}^2}\Big] \;\le\; \frac{1}{1-\gamma}\,\mathbb{E}\Big[\frac{\ell_{t,I_t}^2}{q_{t,I_t}}\Big] = \frac{1}{1-\gamma}\sum_{j=1}^{n}\ell_{t,j}^2 \;\le\; \frac{n}{1-\gamma}.$$

So $\mathrm{Regret}_T \le 3\eta nT + \log(n)/\eta = \sqrt{12\,nT\log n}$ at $\eta = \sqrt{\log(n)/(3nT)}$. ■

The price of bandit feedback, again

Full information $\sqrt{T\log n}$; stochastic $\sqrt{nT\log T}$; adversarial $\sqrt{nT\log n}$. The $\sqrt{n}$ is the price of one coordinate per round. Next: it cannot be removed, even with i.i.d. rewards.

Why $\sqrt{nT}$ cannot be beaten

Two bandits that look the same to any policy

Can anything beat $\sqrt{nT}$?

Experts: $\sqrt{T \log n}$. Bandits: elimination, UCB, Thompson, EXP3 all give $\sqrt{nT}$ up to logs. Weak proofs, or a real price?

Theorem (notes Thm. 18; Lattimore–Szepesvári Thm. 15.2)

Let $T \geq n \geq 2$. For *every* policy π there is an instance with Gaussian rewards and means $\theta^* \in [0,1]^n$ such that

$$R_T(\pi, \theta^*) \;\geq\; c\sqrt{(n-1)\,T}$$

for a universal constant c (e.g. $c = 1/27$).

Can anything beat $\sqrt{nT}$?

Experts: $\sqrt{T \log n}$. Bandits: elimination, UCB, Thompson, EXP3 all give $\sqrt{nT}$ up to logs. Weak proofs, or a real price?

Theorem (notes Thm. 18; Lattimore–Szepesvári Thm. 15.2)

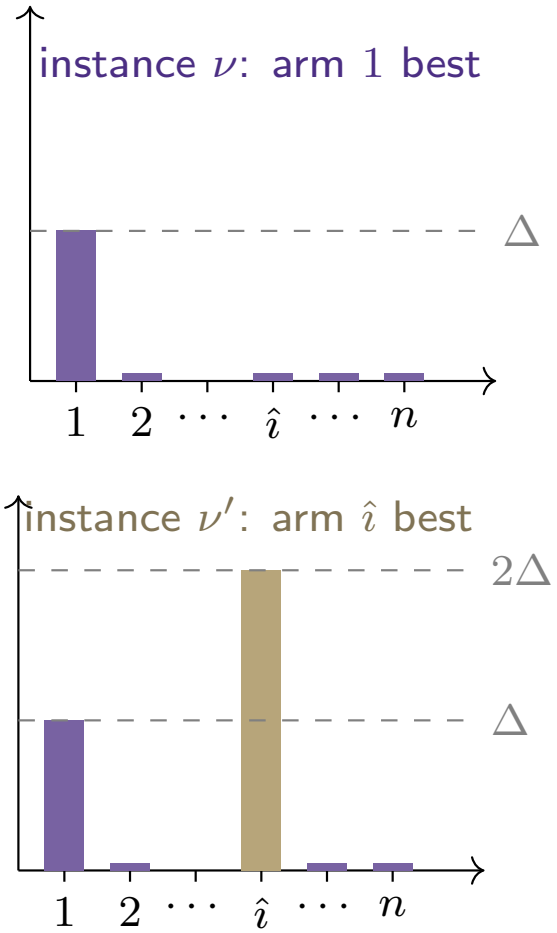
Let $T \geq n \geq 2$. For *every* policy π there is an instance with Gaussian rewards and means $\theta^* \in [0,1]^n$ such that

$$R_T(\pi, \theta^*) \;\geq\; c\sqrt{(n-1)\,T}$$

for a universal constant c (e.g. $c = 1/27$).

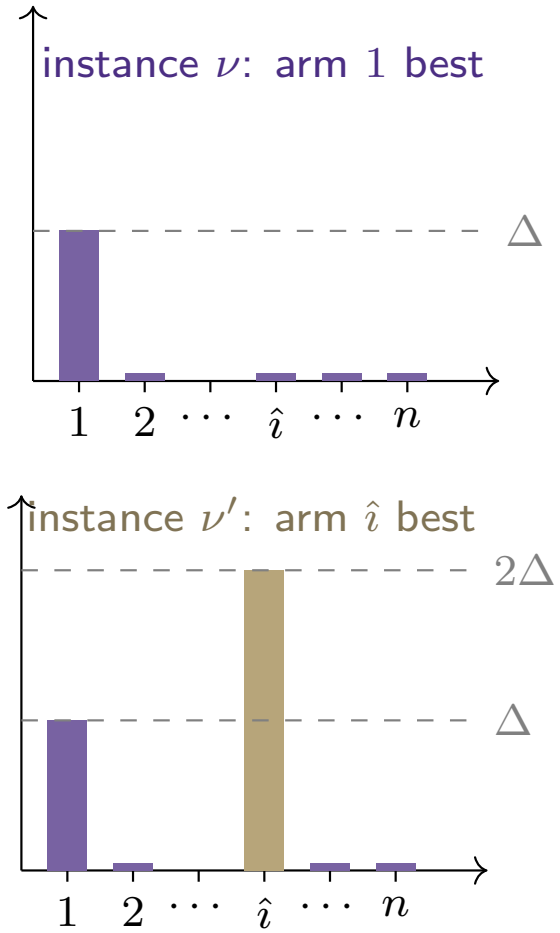
- The quantifiers matter: the hard instance is built *after* seeing the policy. No single instance is hard for every policy.
- The proof is a construction. We will go through the construction and the idea; the two technical facts it rests on we state and do not prove (notes §10.4, Lattimore–Szepesvári Ch. 14–15).

The construction: a bump where the policy is not looking



Step 1. Fix the policy π. Instance ν: $\theta^* = (\Delta, 0, \ldots, 0)$. Run π on ν (in your head) and record how often it pulls each arm.

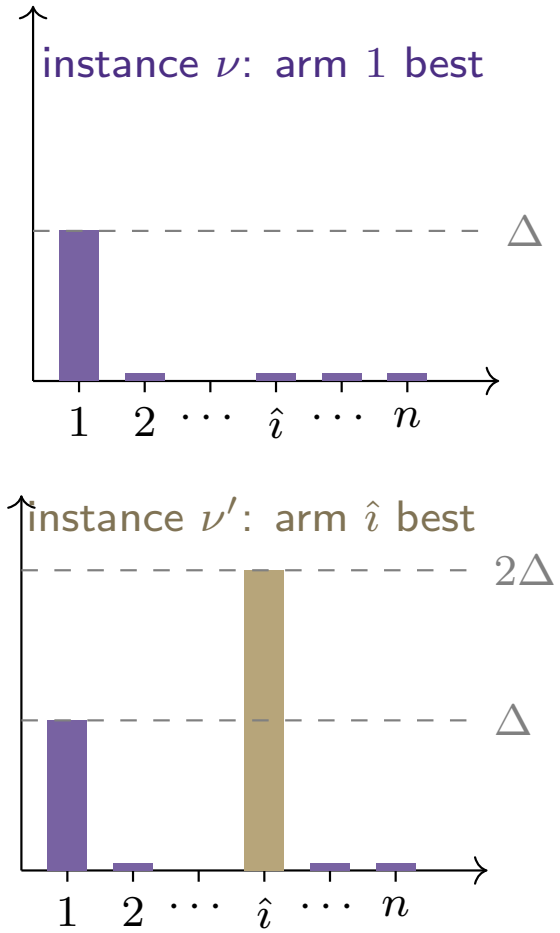
The construction: a bump where the policy is not looking



Step 1. Fix the policy π. Instance ν: $\theta^* = (\Delta, 0, \ldots, 0)$. Run π on ν (in your head) and record how often it pulls each arm.

Step 2. Some arm $\hat{\imath} \neq 1$ is pulled least. Pigeonhole:
$\mathbb{E}_\nu[T_{\hat{\imath}}] \leq$

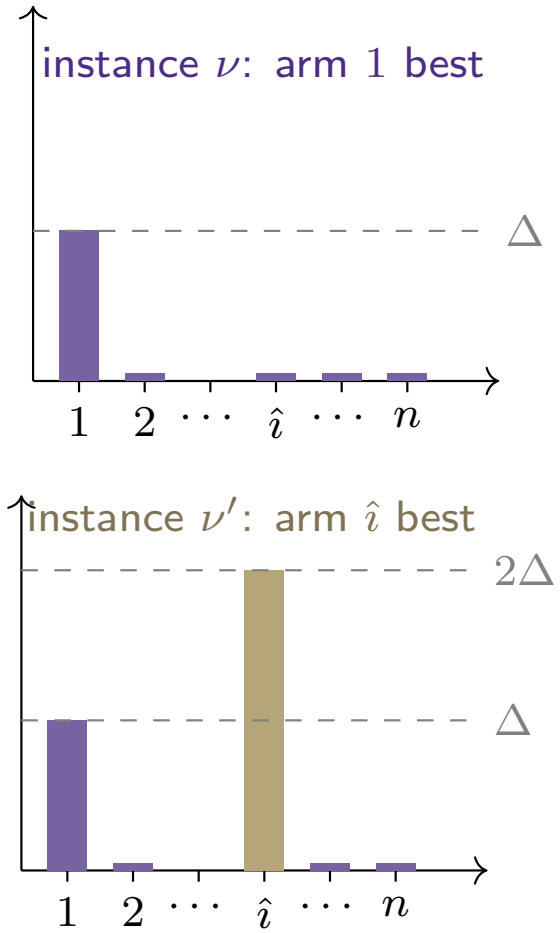
The construction: a bump where the policy is not looking



Step 1. Fix the policy π. Instance ν: $\theta^* = (\Delta, 0, \ldots, 0)$. Run π on ν (in your head) and record how often it pulls each arm.

Step 2. Some arm $\hat{\imath} \neq 1$ is pulled least. Pigeonhole: $\mathbb{E}_\nu[T_{\hat{\imath}}] \leq \frac{1}{n-1} \sum_{i \geq 2} \mathbb{E}_\nu[T_i] \leq \frac{T}{n-1}$.

The construction: a bump where the policy is not looking

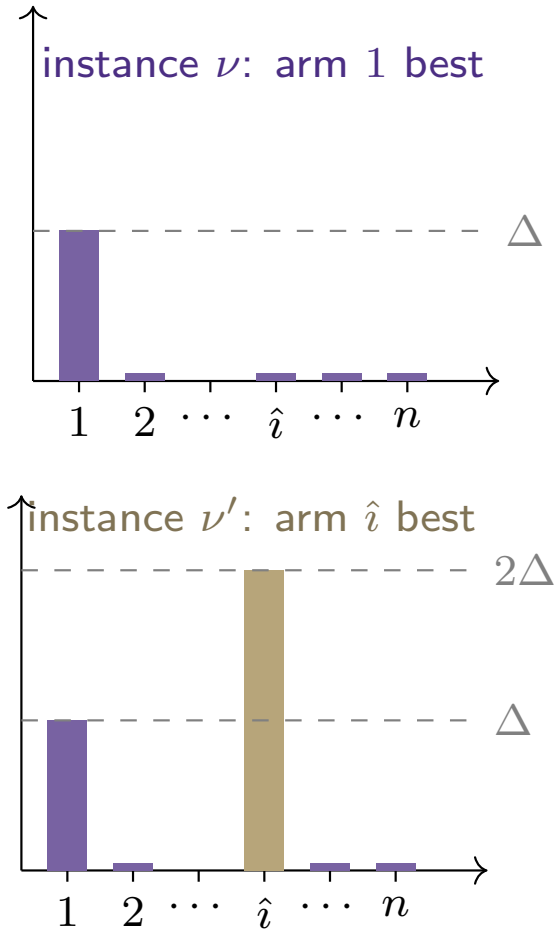


Step 1. Fix the policy π. Instance ν: $\theta^* = (\Delta, 0, \ldots, 0)$. Run π on ν (in your head) and record how often it pulls each arm.

Step 2. Some arm $\hat{\imath} \neq 1$ is pulled least. Pigeonhole: $\mathbb{E}_\nu[T_{\hat{\imath}}] \leq \frac{1}{n-1} \sum_{i \geq 2} \mathbb{E}_\nu[T_i] \leq \frac{T}{n-1}$.

Step 3. Instance ν': same as ν, but raise arm $\hat{\imath}$ to 2Δ. Now $\hat{\imath}$ is best and arm 1 is Δ behind.

The construction: a bump where the policy is not looking



Step 1. Fix the policy π. Instance ν: $\theta^* = (\Delta, 0, \ldots, 0)$. Run π on ν (in your head) and record how often it pulls each arm.

Step 2. Some arm $\hat{\imath} \neq 1$ is pulled least. Pigeonhole: $\mathbb{E}_\nu[T_{\hat{\imath}}] \leq \frac{1}{n-1} \sum_{i \geq 2} \mathbb{E}_\nu[T_i] \leq \frac{T}{n-1}$.

Step 3. Instance ν': same as ν, but raise arm $\hat{\imath}$ to 2Δ. Now $\hat{\imath}$ is best and arm 1 is Δ behind.

The point. The only arm that changed is the one π looks at least: about $T/(n-1)$ samples. The standard error of its mean is $\sqrt{(n-1)/T}$. Choose $\Delta = \frac{1}{2}\sqrt{(n-1)/T}$: the bump 2Δ is one standard error. π cannot tell ν from ν'.

Why the policy pays on one of them

Let $E = \{T_1 \geq T/2\}$: arm 1 gets at least half the pulls. Under ν every pull of an arm $i \neq 1$ costs Δ; under ν' every pull of arm 1 costs Δ:

$$R_T(\pi, \nu) \;\geq \qquad\qquad R_T(\pi, \nu') \;\geq$$

Why the policy pays on one of them

Let $E = \{T_1 \geq T/2\}$: arm 1 gets at least half the pulls. Under ν every pull of an arm $i \neq 1$ costs Δ; under ν' every pull of arm 1 costs Δ:

$$R_T(\pi, \nu) \;\geq\; \Delta \cdot \frac{T}{2} \cdot \mathbb{P}_\nu(E^c), \qquad\qquad R_T(\pi, \nu') \;\geq\; \Delta \cdot \frac{T}{2} \cdot \mathbb{P}_{\nu'}(E).$$

Why the policy pays on one of them

Let $E = \{T_1 \geq T/2\}$: arm 1 gets at least half the pulls. Under ν every pull of an arm $i \neq 1$ costs Δ; under ν' every pull of arm 1 costs Δ:

$$R_T(\pi, \nu) \;\geq\; \Delta \cdot \frac{T}{2} \cdot \mathbb{P}_\nu(E^c), \qquad\qquad R_T(\pi, \nu') \;\geq\; \Delta \cdot \frac{T}{2} \cdot \mathbb{P}_{\nu'}(E).$$

Whatever π does, it either plays arm 1 a lot (bad on ν') or not (bad on ν). If the two instances were literally the same, $\mathbb{P}_\nu(E^c) + \mathbb{P}_{\nu'}(E)$ would be 1. They are only *nearly* the same:

$$\begin{aligned} R_T(\pi, \nu) + R_T(\pi, \nu') \;&\geq\; \frac{\Delta T}{2}\Big[\mathbb{P}_\nu(E^c) + \mathbb{P}_{\nu'}(E)\Big] \\ &\geq \end{aligned}$$

Why the policy pays on one of them

Let $E = \{T_1 \geq T/2\}$: arm 1 gets at least half the pulls. Under ν every pull of an arm $i \neq 1$ costs Δ; under ν' every pull of arm 1 costs Δ:

$$R_T(\pi, \nu) \;\geq\; \Delta \cdot \frac{T}{2} \cdot \mathbb{P}_\nu(E^c), \qquad\qquad R_T(\pi, \nu') \;\geq\; \Delta \cdot \frac{T}{2} \cdot \mathbb{P}_{\nu'}(E).$$

Whatever π does, it either plays arm 1 a lot (bad on ν') or not (bad on ν). If the two instances were literally the same, $\mathbb{P}_\nu(E^c) + \mathbb{P}_{\nu'}(E)$ would be 1. They are only *nearly* the same:

$$\begin{aligned} R_T(\pi, \nu) + R_T(\pi, \nu') \;&\geq\; \frac{\Delta T}{2}\Big[\mathbb{P}_\nu(E^c) + \mathbb{P}_{\nu'}(E)\Big] \\ &\geq\; \frac{\Delta T}{2} \cdot \frac{e^{-1/2}}{2} \;=\; \frac{e^{-1/2}}{8}\sqrt{(n-1)T} \;\geq\; \frac{2}{27}\sqrt{(n-1)T}, \end{aligned}$$

provided $\mathbb{P}_\nu(E^c) + \mathbb{P}_{\nu'}(E) \geq \frac{1}{2}e^{-1/2}$: no test can reject both instances. Then one of the two regrets is at least $\frac{1}{27}\sqrt{(n-1)T}$.

Why the policy pays on one of them

Let $E = \{T_1 \geq T/2\}$: arm 1 gets at least half the pulls. Under ν every pull of an arm $i \neq 1$ costs Δ; under ν' every pull of arm 1 costs Δ:

$$R_T(\pi, \nu) \;\geq\; \Delta \cdot \frac{T}{2} \cdot \mathbb{P}_\nu(E^c), \qquad\qquad R_T(\pi, \nu') \;\geq\; \Delta \cdot \frac{T}{2} \cdot \mathbb{P}_{\nu'}(E).$$

Whatever π does, it either plays arm 1 a lot (bad on ν') or not (bad on ν). If the two instances were literally the same, $\mathbb{P}_\nu(E^c) + \mathbb{P}_{\nu'}(E)$ would be 1. They are only *nearly* the same:

$$\begin{aligned} R_T(\pi, \nu) + R_T(\pi, \nu') \;&\geq\; \frac{\Delta T}{2}\Big[\mathbb{P}_\nu(E^c) + \mathbb{P}_{\nu'}(E)\Big] \\ &\geq\; \frac{\Delta T}{2} \cdot \frac{e^{-1/2}}{2} \;=\; \frac{e^{-1/2}}{8}\sqrt{(n-1)T} \;\geq\; \frac{2}{27}\sqrt{(n-1)T}, \end{aligned}$$

provided $\mathbb{P}_\nu(E^c) + \mathbb{P}_{\nu'}(E) \geq \frac{1}{2}e^{-1/2}$: no test can reject both instances. Then one of the two regrets is at least $\frac{1}{27}\sqrt{(n-1)T}$.

What is left to believe

"π cannot tell ν from ν'" means: for every event E about the transcript, $\mathbb{P}_\nu(E^c) + \mathbb{P}_{\nu'}(E) \geq \frac{1}{2}e^{-1/2}$. Next slide.

Making "cannot tell them apart" precise, in two facts

$\mathbb{P}_\nu, \mathbb{P}_{\nu'}$: the laws of the transcript $(I_1, X_1, \ldots, I_T, X_T)$ when π plays ν or ν'.

Fact 1: information comes only from arms you pull (notes Lemma 17)

$$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{i=1}^{n} \mathbb{E}_\nu[T_i]\, \mathrm{KL}\big(\mathcal{N}(\theta_i, 1), \mathcal{N}(\theta_i', 1)\big) = \sum_{i=1}^{n} \mathbb{E}_\nu[T_i]\, \frac{(\theta_i - \theta_i')^2}{2}.$$

Making "cannot tell them apart" precise, in two facts

$\mathbb{P}_\nu, \mathbb{P}_{\nu'}$: the laws of the transcript $(I_1, X_1, \ldots, I_T, X_T)$ when π plays ν or ν'.

Fact 1: information comes only from arms you pull (notes Lemma 17)

$$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{i=1}^{n} \mathbb{E}_\nu[T_i]\, \mathrm{KL}\big(\mathcal{N}(\theta_i, 1), \mathcal{N}(\theta'_i, 1)\big) = \sum_{i=1}^{n} \mathbb{E}_\nu[T_i]\, \frac{(\theta_i - \theta'_i)^2}{2}.$$

Here only arm $\hat{\imath}$ differs, by 2Δ, and it is pulled $\le T/(n-1)$ times:
$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \le$

Making "cannot tell them apart" precise, in two facts

$\mathbb{P}_\nu, \mathbb{P}_{\nu'}$: the laws of the transcript $(I_1, X_1, \ldots, I_T, X_T)$ when π plays ν or ν'.

Fact 1: information comes only from arms you pull (notes Lemma 17)

$$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{i=1}^{n} \mathbb{E}_\nu[T_i]\, \mathrm{KL}\big(\mathcal{N}(\theta_i, 1), \mathcal{N}(\theta_i', 1)\big) = \sum_{i=1}^{n} \mathbb{E}_\nu[T_i]\, \frac{(\theta_i - \theta_i')^2}{2}.$$

Here only arm $\hat{\imath}$ differs, by 2Δ, and it is pulled $\le T/(n-1)$ times:
$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \le \frac{T}{n-1} \cdot \frac{(2\Delta)^2}{2} = \frac{2\Delta^2 T}{n-1} = \frac{1}{2}$ at $\Delta = \frac{1}{2}\sqrt{(n-1)/T}$.

Making "cannot tell them apart" precise, in two facts

$\mathbb{P}_\nu, \mathbb{P}_{\nu'}$: the laws of the transcript $(I_1, X_1, \ldots, I_T, X_T)$ when π plays ν or ν'.

Fact 1: information comes only from arms you pull (notes Lemma 17)

$$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{i=1}^{n} \mathbb{E}_\nu[T_i]\, \mathrm{KL}\big(\mathcal{N}(\theta_i, 1), \mathcal{N}(\theta'_i, 1)\big) = \sum_{i=1}^{n} \mathbb{E}_\nu[T_i]\, \frac{(\theta_i - \theta'_i)^2}{2}.$$

Here only arm $\hat{\imath}$ differs, by 2Δ, and it is pulled $\le T/(n-1)$ times:
$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \le \frac{T}{n-1} \cdot \frac{(2\Delta)^2}{2} = \frac{2\Delta^2 T}{n-1} = \frac{1}{2}$ at $\Delta = \frac{1}{2}\sqrt{(n-1)/T}$.

Fact 2: small KL means no event can separate them (Bretagnolle–Huber inequality)

For any event E: $\quad \mathbb{P}_\nu(E^c) + \mathbb{P}_{\nu'}(E) \;\ge\; \frac{1}{2}\exp\big(-\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'})\big) \;\ge\; \frac{1}{2}e^{-1/2}$.

Exactly what the previous slide needed. ■ We will use the same two facts for both instance-dependent bounds.

Making "cannot tell them apart" precise, in two facts

$\mathbb{P}_\nu, \mathbb{P}_{\nu'}$: the laws of the transcript $(I_1, X_1, \ldots, I_T, X_T)$ when π plays ν or ν'.

Fact 1: information comes only from arms you pull (notes Lemma 17)

$$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{i=1}^{n} \mathbb{E}_\nu[T_i]\, \mathrm{KL}\big(\mathcal{N}(\theta_i, 1), \mathcal{N}(\theta'_i, 1)\big) = \sum_{i=1}^{n} \mathbb{E}_\nu[T_i]\, \frac{(\theta_i - \theta'_i)^2}{2}.$$

Here only arm $\hat{\imath}$ differs, by 2Δ, and it is pulled $\le T/(n-1)$ times:
$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \le \frac{T}{n-1} \cdot \frac{(2\Delta)^2}{2} = \frac{2\Delta^2 T}{n-1} = \frac{1}{2}$ at $\Delta = \frac{1}{2}\sqrt{(n-1)/T}$.

Fact 2: small KL means no event can separate them (Bretagnolle–Huber inequality)

For any event E: $\quad \mathbb{P}_\nu(E^c) + \mathbb{P}_{\nu'}(E) \;\ge\; \frac{1}{2}\exp\big(-\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'})\big) \;\ge\; \frac{1}{2}e^{-1/2}$.

Exactly what the previous slide needed. ■ We will use the same two facts for both instance-dependent bounds.

Where n enters: $T/(n-1)$ samples of the changed arm. **Where $\sqrt{T}$ enters:** a bump $\Delta \approx \sqrt{n/T}$ costs Δ per round for T rounds.

Instance-dependent lower bounds

First best-arm identification, then regret

Pure exploration: the definition of PAC does all the work

An identification algorithm is a sampling rule I_t, a stopping time τ, and a recommendation $\hat{\imath}$ made at time τ.

Definition (δ-PAC)

An algorithm is δ-PAC on a class of instances if for every instance θ^* in the class, $\mathbb{P}_{\theta^*}\left(\hat{\imath} = \arg\max_i \theta_i^*\right) \geq 1 - \delta$.

- The quantifier is the whole point. "Output arm 1 without sampling" is right on our instance but not δ-PAC: it fails where arm 1 is not best. A δ-PAC algorithm must be right on *every* instance, including ones it never sees.
- Successive elimination is δ-PAC with $\mathbb{E}[\tau] \lesssim \sum_{i \geq 2} \Delta_i^{-2} \log\left(n \log(\Delta_i^{-2})/\delta\right)$. Is Δ_i^{-2} per arm necessary?

Pure exploration lower bound: the argument

Theorem (Mannor & Tsitsiklis 2004; Kaufmann et al. 2016; notes Thm. 17)

Any δ-PAC algorithm on Gaussian instances with a unique best arm satisfies, on every such instance, $\mathbb{E}[\tau] \geq \sum_{i \geq 2} 2\Delta_i^{-2} \log(1/(4\delta))$ (sharper tools: $\log(1/(2.4\delta))$).

Pure exploration lower bound: the argument

Theorem (Mannor & Tsitsiklis 2004; Kaufmann et al. 2016; notes Thm. 17)

Any δ-PAC algorithm on Gaussian instances with a unique best arm satisfies, on every such instance, $\mathbb{E}[\tau] \geq \sum_{i \geq 2} 2\Delta_i^{-2} \log(1/(4\delta))$ (sharper tools: $\log(1/(2.4\delta))$).

Construction. Fix the instance ν (arm 1 best) and a suboptimal arm j. Instance ν': identical, except arm j is raised to $\theta_1^* + \epsilon$, so j is best in ν'. Let $E = \{\hat{\imath} = 1\}$.

Pure exploration lower bound: the argument

Theorem (Mannor & Tsitsiklis 2004; Kaufmann et al. 2016; notes Thm. 17)

Any δ-PAC algorithm on Gaussian instances with a unique best arm satisfies, on every such instance, $\mathbb{E}[\tau] \geq \sum_{i \geq 2} 2\Delta_i^{-2} \log(1/(4\delta))$ (sharper tools: $\log(1/(2.4\delta))$).

Construction. Fix the instance ν (arm 1 best) and a suboptimal arm j. Instance ν': identical, except arm j is raised to $\theta_1^* + \epsilon$, so j is best in ν'. Let $E = \{\hat{\imath} = 1\}$.

PAC on both instances. Because the algorithm is δ-PAC on ν *and* on ν':

$\mathbb{P}_\nu(E^c) \leq$ $\qquad$ $\mathbb{P}_{\nu'}(E) \leq$

Pure exploration lower bound: the argument

Theorem (Mannor & Tsitsiklis 2004; Kaufmann et al. 2016; notes Thm. 17)

Any δ-PAC algorithm on Gaussian instances with a unique best arm satisfies, on every such instance, $\mathbb{E}[\tau] \geq \sum_{i \geq 2} 2\Delta_i^{-2} \log(1/(4\delta))$ (sharper tools: $\log(1/(2.4\delta))$).

Construction. Fix the instance ν (arm 1 best) and a suboptimal arm j. Instance ν': identical, except arm j is raised to $\theta_1^* + \epsilon$, so j is best in ν'. Let $E = \{\hat{\imath} = 1\}$.

PAC on both instances. Because the algorithm is δ-PAC on ν *and* on ν': $\mathbb{P}_\nu(E^c) \leq \delta, \qquad \mathbb{P}_{\nu'}(E) \leq \delta.$

Pure exploration lower bound: the argument

Theorem (Mannor & Tsitsiklis 2004; Kaufmann et al. 2016; notes Thm. 17)

Any δ-PAC algorithm on Gaussian instances with a unique best arm satisfies, on every such instance, $\mathbb{E}[\tau] \geq \sum_{i \geq 2} 2\Delta_i^{-2} \log(1/(4\delta))$ (sharper tools: $\log(1/(2.4\delta))$).

Construction. Fix the instance ν (arm 1 best) and a suboptimal arm j. Instance ν': identical, except arm j is raised to $\theta_1^* + \epsilon$, so j is best in ν'. Let $E = \{\hat{\imath} = 1\}$.

PAC on both instances. Because the algorithm is δ-PAC on ν *and* on ν': $\mathbb{P}_\nu(E^c) \leq \delta, \qquad \mathbb{P}_{\nu'}(E) \leq \delta.$

So the transcripts must differ. Bretagnolle–Huber: $2\delta \;\geq\; \mathbb{P}_\nu(E^c) + \mathbb{P}_{\nu'}(E) \;\geq$

Pure exploration lower bound: the argument

Theorem (Mannor & Tsitsiklis 2004; Kaufmann et al. 2016; notes Thm. 17)

Any δ-PAC algorithm on Gaussian instances with a unique best arm satisfies, on every such instance, $\mathbb{E}[\tau] \geq \sum_{i \geq 2} 2\Delta_i^{-2} \log(1/(4\delta))$ (sharper tools: $\log(1/(2.4\delta))$).

Construction. Fix the instance ν (arm 1 best) and a suboptimal arm j. Instance ν': identical, except arm j is raised to $\theta_1^* + \epsilon$, so j is best in ν'. Let $E = \{\hat{\imath} = 1\}$.

PAC on both instances. Because the algorithm is δ-PAC on ν *and* on ν':
$\mathbb{P}_\nu(E^c) \leq \delta, \qquad \mathbb{P}_{\nu'}(E) \leq \delta.$

So the transcripts must differ. Bretagnolle–Huber:
$2\delta \;\geq\; \mathbb{P}_\nu(E^c) + \mathbb{P}_{\nu'}(E) \;\geq\; \frac{1}{2} e^{-\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'})} \qquad \Longrightarrow \qquad \mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \geq \log \frac{1}{4\delta}.$

Pure exploration lower bound: the argument

Theorem (Mannor & Tsitsiklis 2004; Kaufmann et al. 2016; notes Thm. 17)

Any δ-PAC algorithm on Gaussian instances with a unique best arm satisfies, on every such instance, $\mathbb{E}[\tau] \geq \sum_{i \geq 2} 2\Delta_i^{-2} \log(1/(4\delta))$ (sharper tools: $\log(1/(2.4\delta))$).

Construction. Fix the instance ν (arm 1 best) and a suboptimal arm j. Instance ν': identical, except arm j is raised to $\theta_1^* + \epsilon$, so j is best in ν'. Let $E = \{\hat{\imath} = 1\}$.

PAC on both instances. Because the algorithm is δ-PAC on ν *and* on ν':
$\mathbb{P}_\nu(E^c) \leq \delta, \qquad \mathbb{P}_{\nu'}(E) \leq \delta.$

So the transcripts must differ. Bretagnolle–Huber:
$2\delta \;\geq\; \mathbb{P}_\nu(E^c) + \mathbb{P}_{\nu'}(E) \;\geq\; \frac{1}{2} e^{-\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'})} \qquad \Longrightarrow \qquad \mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \geq \log \frac{1}{4\delta}.$

In words: a δ-PAC algorithm *must* make the transcripts differ by $\geq \log(1/(4\delta))$ nats of KL. Where can that come from?

Where the information comes from: arm j

Fact 1 (information adds up over pulls) also holds at a stopping time. Only arm j differs between ν and ν', by $\Delta_j + \epsilon$, so

$$\log \frac{1}{4\delta} \;\leq\; \mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \;=$$

Where the information comes from: arm j

Fact 1 (information adds up over pulls) also holds at a stopping time. Only arm j differs between ν and ν', by $\Delta_j + \epsilon$, so

$$\log \frac{1}{4\delta} \;\le\; \mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \;=\; \mathbb{E}_\nu[T_j]\, \frac{(\Delta_j + \epsilon)^2}{2},$$

so $\mathbb{E}_\nu[T_j] \ge 2\log(1/(4\delta))/(\Delta_j + \epsilon)^2$.

Where the information comes from: arm j

Fact 1 (information adds up over pulls) also holds at a stopping time. Only arm j differs between ν and ν', by $\Delta_j + \epsilon$, so

$$\log \frac{1}{4\delta} \;\le\; \mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \;=\; \mathbb{E}_\nu[T_j]\, \frac{(\Delta_j + \epsilon)^2}{2},$$

so $\mathbb{E}_\nu[T_j] \ge 2\log(1/(4\delta))/(\Delta_j + \epsilon)^2$. The arm j and the ϵ were arbitrary; let $\epsilon \to 0$ and sum over j:

$$\mathbb{E}[\tau] \;=\; \sum_{j=1}^{n} \mathbb{E}[T_j] \;\ge\; \sum_{j \ge 2} \frac{2\log(1/(4\delta))}{\Delta_j^2}. \qquad \blacksquare$$

Where the information comes from: arm j

Fact 1 (information adds up over pulls) also holds at a stopping time. Only arm j differs between ν and ν', by $\Delta_j + \epsilon$, so

$$\log \frac{1}{4\delta} \;\le\; \mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \;=\; \mathbb{E}_\nu[T_j]\, \frac{(\Delta_j + \epsilon)^2}{2},$$

so $\mathbb{E}_\nu[T_j] \ge 2\log(1/(4\delta))/(\Delta_j + \epsilon)^2$. The arm j and the ϵ were arbitrary; let $\epsilon \to 0$ and sum over j:

$$\mathbb{E}[\tau] \;=\; \sum_{j=1}^{n} \mathbb{E}[T_j] \;\ge\; \sum_{j \ge 2} \frac{2\log(1/(4\delta))}{\Delta_j^2}. \qquad \blacksquare$$

What made this easy

- The δ-PAC definition quantifies over *all* instances, so it tells us what the algorithm does on ν': it must recommend j.
- Each arm gets its own alternative instance, so *every* Δ_j^{-2} appears, not only the smallest gap.
- Regret makes no promise about ν'. We will *assume* one (consistency); then the same argument goes through.

Instance by instance: $\log T/\Delta_i$ is also unavoidable

Now regret. On an easy instance UCB gets $\sum_i \log T/\Delta_i \ll \sqrt{nT}$. Is that right? The stand-in for "δ-PAC on every instance" is:

Theorem (Lai & Robbins 1985; notes Lemma 18)

Fix a Gaussian instance and a consistent policy: $R_T = o(T^a)$ on every instance, for every $a > 0$. Then for every arm with $\Delta_i > 0$,
$\mathbb{E}[T_i] \geq (2 - o(1))\,\Delta_i^{-2} \log T$, hence $R_T \geq (1 - o(1)) \sum_{i \geq 2} 2 \log T/\Delta_i$.

Instance by instance: $\log T/\Delta_i$ is also unavoidable

Now regret. On an easy instance UCB gets $\sum_i \log T/\Delta_i \ll \sqrt{nT}$. Is that right? The stand-in for "δ-PAC on every instance" is:

Theorem (Lai & Robbins 1985; notes Lemma 18)

Fix a Gaussian instance and a consistent policy: $R_T = o(T^a)$ on every instance, for every $a > 0$. Then for every arm with $\Delta_i > 0$,
$\mathbb{E}[T_i] \geq (2 - o(1))\,\Delta_i^{-2} \log T$, hence $R_T \geq (1 - o(1)) \sum_{i \geq 2} 2 \log T/\Delta_i$.

Same construction as for PAC. Fix the instance ν and a suboptimal arm i. Instance ν': identical, except arm i is raised to $\theta_1^* + \epsilon$. Consistency on ν' plays the role that PAC on ν' played:

- Under ν a consistent policy plays arm i rarely:
  $\mathbb{P}_\nu(T_i \geq T/2) \leq 2\mathbb{E}_\nu[T_i]/T = o(T^{a-1})$.
- Under ν' arm i is best, so it plays i almost always:
  $\mathbb{P}_{\nu'}(T_i < T/2) \leq 2\mathbb{E}_{\nu'}[T - T_i]/T = o(T^{a-1})$.
- So the policy *does* tell ν from ν', with polynomially small error. That costs information, which only comes from pulling arm i.

The calculation, in one line

Telling two instances apart with error $\approx T^{-a}$ needs KL of order $\log T$. Precisely: Bretagnolle–Huber gives $\mathbb{P}_\nu(E) + \mathbb{P}_{\nu'}(E^c) \geq \frac{1}{2} e^{-\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'})}$, and with $E = \{T_i \geq T/2\}$ the left side is $o(T^{a-1})$. Hence

$$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \;\geq$$

The calculation, in one line

Telling two instances apart with error $\approx T^{-a}$ needs KL of order $\log T$. Precisely: Bretagnolle–Huber gives $\mathbb{P}_\nu(E) + \mathbb{P}_{\nu'}(E^c) \geq \frac{1}{2} e^{-\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'})}$, and with $E = \{T_i \geq T/2\}$ the left side is $o(T^{a-1})$. Hence

$$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \;\geq\; (1-a)\log T - O(1).$$

The calculation, in one line

Telling two instances apart with error $\approx T^{-a}$ needs KL of order $\log T$. Precisely: Bretagnolle–Huber gives $\mathbb{P}_\nu(E) + \mathbb{P}_{\nu'}(E^c) \geq \frac{1}{2} e^{-\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'})}$, and with $E = \{T_i \geq T/2\}$ the left side is $o(T^{a-1})$. Hence

$$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \;\geq\; (1-a)\log T - O(1).$$

Meanwhile only arm i differs between the instances, by $\Delta_i + \epsilon$, so Fact 1 (information adds up over pulls) gives

$$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \;=$$

The calculation, in one line

Telling two instances apart with error $\approx T^{-a}$ needs KL of order $\log T$. Precisely: Bretagnolle–Huber gives $\mathbb{P}_\nu(E) + \mathbb{P}_{\nu'}(E^c) \geq \frac{1}{2} e^{-\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'})}$, and with $E = \{T_i \geq T/2\}$ the left side is $o(T^{a-1})$. Hence

$$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \;\geq\; (1-a)\log T - O(1).$$

Meanwhile only arm i differs between the instances, by $\Delta_i + \epsilon$, so Fact 1 (information adds up over pulls) gives

$$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \;=\; \mathbb{E}_\nu[T_i]\,\frac{(\Delta_i+\epsilon)^2}{2}.$$

The calculation, in one line

Telling two instances apart with error $\approx T^{-a}$ needs KL of order $\log T$. Precisely: Bretagnolle–Huber gives $\mathbb{P}_\nu(E) + \mathbb{P}_{\nu'}(E^c) \geq \frac{1}{2} e^{-\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'})}$, and with $E = \{T_i \geq T/2\}$ the left side is $o(T^{a-1})$. Hence

$$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \;\geq\; (1-a)\log T - O(1).$$

Meanwhile only arm i differs between the instances, by $\Delta_i + \epsilon$, so Fact 1 (information adds up over pulls) gives

$$\mathrm{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \;=\; \mathbb{E}_\nu[T_i]\,\frac{(\Delta_i+\epsilon)^2}{2}.$$

Combine, then let $a \to 0$ and $\epsilon \to 0$:

$$\mathbb{E}_\nu[T_i] \;\geq\; \frac{2(1-a)\log T - O(1)}{(\Delta_i+\epsilon)^2} \;\longrightarrow\; \frac{2\log T}{\Delta_i^2}. \qquad \blacksquare$$

Each pull of arm i costs Δ_i, so $R_T \gtrsim \sum_i 2\log T/\Delta_i$: the price of being *sure* arm i is not the best one.

Reading the two lower bounds together

- **Minimax, $\sqrt{nT}$:** the hard instance has gaps $\approx \sqrt{n/T}$, and the policy *never* finds out which arm was bumped.
- **Instance-dependent, $\sum_i 2\log T/\Delta_i$:** on a fixed instance a good policy *does* find out, and $2\log T/\Delta_i^2$ pulls of arm i is the cost of being sure. This one comes with the constant.
- **Who matches them.** UCB with width $\sqrt{2\alpha \log t / T_i(t)}$, $\alpha \downarrow 1$, and Thompson sampling hit the constant 2; elimination does not (it keeps pulling every survivor). MOSS hits $\sqrt{nT}$ exactly; Lattimore (2018) gets both at once. EXP3's $\sqrt{nT\log n}$ is tight up to the log, with no distributional assumption.
- **One template.** Two instances that differ only where the policy looks least; a KL that adds up over the pulls; a change-of-measure inequality (Bretagnolle–Huber, all three times). We will reuse it for linear bandits and active learning.