

CSE 541: Interactive Learning

Lecture 7: UCB, Thompson sampling, EXP3, and a lower bound

Kevin Jamieson

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

Wednesday, October 21, 2026



Last time: successive elimination

n arms, 1-sub-Gaussian rewards with means θ_i^* , gaps $\Delta_i = \max_j \theta_j^* - \theta_i^*$, T_i pulls of arm i . Keep a set of survivors, pull them equally, and at geometrically increasing precision drop any arm confidently worse than the leader.

Successive elimination (notes, §10.2)

Input: arms $\mathcal{X} = [n]$, confidence $\delta \in (0, 1)$. **Initialize** $\mathcal{X}_1 = \mathcal{X}$, $\ell = 1$.

while $|\mathcal{X}_\ell| > 1$:

$\epsilon_\ell = 2^{-\ell}$; pull each arm in $\mathcal{X}_\ell$ exactly $\tau_\ell = \lceil 2\epsilon_\ell^{-2} \log(4\ell^2 n/\delta) \rceil$ times

 compute the empirical means $\hat{\theta}_{i,\ell}$ of these τ_ℓ rewards, $i \in \mathcal{X}_\ell$

$\mathcal{X}_{\ell+1} = \mathcal{X}_\ell \setminus \{i \in \mathcal{X}_\ell : \max_{j \in \mathcal{X}_\ell} \hat{\theta}_{j,\ell} - \hat{\theta}_{i,\ell} > 2\epsilon_\ell\}$; $\ell \leftarrow \ell + 1$

Output the surviving arm (for regret: keep playing it forever).

- τ_ℓ makes each round- ℓ empirical mean ϵ_ℓ -accurate w.p. $\geq 1 - \delta/(2n\ell^2)$; the ℓ^2 pays for a union bound over all rounds.
- Sample size quadruples each round, so the last round dominates. No Δ_i appears anywhere.

Last time: what successive elimination guarantees

Good event $E = \bigcap_{i,\ell} \{|\hat{\theta}_{i,\ell} - \theta_i^*| \leq \epsilon_\ell\}$; assume $\max_i \Delta_i \leq 4$ and arm 1 is best.

Lemma

$\mathbb{P}(E) \geq 1 - \delta$, and on E , for every ℓ : $1 \in \mathcal{X}_\ell$ and $\max_{i \in \mathcal{X}_\ell} \Delta_i \leq 8\epsilon_\ell$.

Theorem (sample complexity)

W.p. $\geq 1 - \delta$, arm i is pulled $T_i \leq c \Delta_i^{-2} \log(n \log(\Delta_i^{-2})/\delta)$ times, so the best arm is found after $c \sum_{i \geq 2} \Delta_i^{-2} \log(\dots)$ pulls.

Theorem (regret)

W.p. $\geq 1 - \delta$, $\Delta_i T_i \lesssim \Delta_i^{-1} L$ for every arm and every T , where $L = \log(n \log T/\delta)$.
With $\delta = 1/T$:

$$R_T \lesssim \sum_{i \geq 2} \frac{\log T}{\Delta_i} \quad \text{and} \quad R_T \leq \sum_i \min \left\{ \Delta_i T_i, \frac{cL}{\Delta_i} \right\} \leq \sqrt{cL n T}.$$

Last time: what successive elimination guarantees

Good event $E = \bigcap_{i,\ell} \{|\hat{\theta}_{i,\ell} - \theta_i^*| \leq \epsilon_\ell\}$; assume $\max_i \Delta_i \leq 4$ and arm 1 is best.

Lemma

$\mathbb{P}(E) \geq 1 - \delta$, and on E , for every ℓ : $1 \in \mathcal{X}_\ell$ and $\max_{i \in \mathcal{X}_\ell} \Delta_i \leq 8\epsilon_\ell$.

Theorem (sample complexity)

W.p. $\geq 1 - \delta$, arm i is pulled $T_i \leq c \Delta_i^{-2} \log(n \log(\Delta_i^{-2})/\delta)$ times, so the best arm is found after $c \sum_{i \geq 2} \Delta_i^{-2} \log(\dots)$ pulls.

Theorem (regret)

W.p. $\geq 1 - \delta$, $\Delta_i T_i \lesssim \Delta_i^{-1} L$ for every arm and every T , where $L = \log(n \log T/\delta)$.
With $\delta = 1/T$:

$$R_T \lesssim \sum_{i \geq 2} \frac{\log T}{\Delta_i} \quad \text{and} \quad R_T \leq \sum_i \min \left\{ \Delta_i T_i, \frac{cL}{\Delta_i} \right\} \leq \sqrt{cL n T}.$$

Today. Two algorithms people actually run (UCB, Thompson sampling); bandits with no i.i.d. assumption (EXP3); and why the $\sqrt{n}$ cannot be removed.

Two algorithms you will actually use

Optimism, and sampling from the posterior

UCB: optimism in the face of uncertainty

Elimination pulls survivors equally and acts only at the end of a round. More natural: at every step, give each arm the benefit of the doubt and play the one that *could* be best.

UCB (notes §10.4.4; Auer, Cesa-Bianchi, Fischer 2002)

Input: $\delta \in (0, 1)$. Pull each arm once.

for $t = n + 1, n + 2, \dots$: with $T_i(t)$ the pulls of arm i so far and $\hat{\theta}_i(t)$ its empirical mean, play

$$I_t = \arg \max_{i \in [n]} \text{UCB}_i(t), \quad \text{UCB}_i(t) := \hat{\theta}_i(t) + \sqrt{\frac{2 \log(2nT^2/\delta)}{T_i(t)}}$$

UCB: optimism in the face of uncertainty

Elimination pulls survivors equally and acts only at the end of a round. More natural: at every step, give each arm the benefit of the doubt and play the one that *could* be best.

UCB (notes §10.4.4; Auer, Cesa-Bianchi, Fischer 2002)

Input: $\delta \in (0, 1)$. Pull each arm once.

for $t = n + 1, n + 2, \dots$: with $T_i(t)$ the pulls of arm i so far and $\hat{\theta}_i(t)$ its empirical mean, play

$$I_t = \arg \max_{i \in [n]} \text{UCB}_i(t), \quad \text{UCB}_i(t) := \hat{\theta}_i(t) + \sqrt{\frac{2 \log(2nT^2/\delta)}{T_i(t)}}$$

- The width is the A/B confidence interval, union-bounded over n arms and $t \leq T$.
- An arm is played because its mean looks high (exploit) or because it is rarely pulled and its interval is wide (explore). No exploration schedule, no Δ_i , no rounds. Each play of a bad arm *shrinks* its interval, so it cannot keep winning the argmax.

UCB: why a bad arm cannot be played often

Let E be the event that every interval is valid:

$|\hat{\theta}_i(t) - \theta_i^*| \leq \sqrt{2 \log(2nT^2/\delta)/T_i(t)}$ for all i and all $t \leq T$; by the union bound $\mathbb{P}(E) \geq 1 - \delta$.

UCB: why a bad arm cannot be played often

Let E be the event that every interval is valid:

$|\hat{\theta}_i(t) - \theta_i^*| \leq \sqrt{2 \log(2nT^2/\delta)/T_i(t)}$ for all i and all $t \leq T$; by the union bound $\mathbb{P}(E) \geq 1 - \delta$.

Suppose arm $i \neq 1$ is played at time t (so $\text{UCB}_i(t) \geq \text{UCB}_1(t)$). On E ,

$$\theta_1^* \leq \text{UCB}_1(t) \leq \text{UCB}_i(t) \leq$$

UCB: why a bad arm cannot be played often

Let E be the event that every interval is valid:

$|\hat{\theta}_i(t) - \theta_i^*| \leq \sqrt{2 \log(2nT^2/\delta)/T_i(t)}$ for all i and all $t \leq T$; by the union bound $\mathbb{P}(E) \geq 1 - \delta$.

Suppose arm $i \neq 1$ is played at time t (so $\text{UCB}_i(t) \geq \text{UCB}_1(t)$). On E ,

$$\theta_1^* \leq \text{UCB}_1(t) \leq \text{UCB}_i(t) \leq \theta_i^* + 2\sqrt{\frac{2 \log(2nT^2/\delta)}{T_i(t)}},$$

so $T_i(t) \leq 8\Delta_i^{-2} \log(2nT^2/\delta)$. This holds the last time arm i is played, so it is played at most $1 + 8\Delta_i^{-2} \log(2nT^2/\delta)$ times in total.

UCB: why a bad arm cannot be played often

Let E be the event that every interval is valid:

$|\hat{\theta}_i(t) - \theta_i^*| \leq \sqrt{2 \log(2nT^2/\delta)/T_i(t)}$ for all i and all $t \leq T$; by the union bound $\mathbb{P}(E) \geq 1 - \delta$.

Suppose arm $i \neq 1$ is played at time t (so $\text{UCB}_i(t) \geq \text{UCB}_1(t)$). On E ,

$$\theta_1^* \leq \text{UCB}_1(t) \leq \text{UCB}_i(t) \leq \theta_i^* + 2\sqrt{\frac{2 \log(2nT^2/\delta)}{T_i(t)}},$$

so $T_i(t) \leq 8\Delta_i^{-2} \log(2nT^2/\delta)$. This holds the last time arm i is played, so it is played at most $1 + 8\Delta_i^{-2} \log(2nT^2/\delta)$ times in total.

Theorem (UCB regret)

With $\delta = 1/T$: $R_T \lesssim \sum_{i \geq 2} \frac{\log(nT)}{\Delta_i}$ and, by the same $\min\{\Delta_i T_i, \log(nT)/\Delta_i\}$ trick as before, $R_T \lesssim \sqrt{nT \log(nT)}$.

Same rates as elimination, but anytime, adaptive, and one line of code. Sharper versions (UCB with $\sqrt{2 \log t/T_i(t)}$, KL-UCB, MOSS) reach the exact asymptotic constant $\sum_i 2 \log T/\Delta_i$ and the minimax $\sqrt{nT}$.

Thompson sampling: optimism by randomization

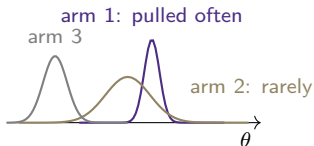
Be Bayesian about the means. With a flat prior and $\mathcal{N}(\theta_i^*, 1)$ rewards, after T_i pulls the posterior of θ_i^* is $\mathcal{N}(\hat{\theta}_i, 1/T_i)$: the same center and the same width as UCB's interval.

Thompson sampling (Thompson 1933)

for $t = 1, 2, \dots$:

sample $\tilde{\theta}_i \sim \text{posterior}_i(t)$ independently for each arm i

play $I_t = \arg \max_i \tilde{\theta}_i$; update the posterior of I_t with the observed reward



UCB plays the arm with the highest upper quantile: deterministic optimism.

Thompson plays arm i with probability equal to the posterior probability that i is the best arm (“probability matching”). A rarely pulled arm has a wide posterior, so its sample is occasionally the largest; once it is pulled, the posterior narrows.

Thompson sampling: what is known, and why people like it

- **Regret.** For Gaussian or Bernoulli rewards, Thompson sampling achieves $R_T \lesssim \sum_{i \geq 2} \log T / \Delta_i$ and $R_T \lesssim \sqrt{nT \log n}$, and in fact the exact asymptotic constant $\sum_i 2 \log T / \Delta_i$ (Agrawal & Goyal 2012, 2013; Kaufmann, Korda & Munos 2012). Same rates as UCB.
- **Proof idea.** Optimism holds with constant probability: the sample for the best arm exceeds its mean a constant fraction of the time, which is enough to make the UCB argument go through in expectation.
- **In practice** it is often better than UCB (Chapelle & Li 2011): union bounds over n and T make UCB's intervals wider than they need to be, while the posterior uses exactly the information available.
- **It generalizes effortlessly.** Any model with a posterior you can sample from (contextual, structured, delayed or batched feedback) gives an algorithm. Guarantees for those settings are harder, and are where the theory is still being written.

Thompson sampling: what is known, and why people like it

- **Regret.** For Gaussian or Bernoulli rewards, Thompson sampling achieves $R_T \lesssim \sum_{i \geq 2} \log T / \Delta_i$ and $R_T \lesssim \sqrt{nT \log n}$, and in fact the exact asymptotic constant $\sum_i 2 \log T / \Delta_i$ (Agrawal & Goyal 2012, 2013; Kaufmann, Korda & Munos 2012). Same rates as UCB.
- **Proof idea.** Optimism holds with constant probability: the sample for the best arm exceeds its mean a constant fraction of the time, which is enough to make the UCB argument go through in expectation.
- **In practice** it is often better than UCB (Chapelle & Li 2011): union bounds over n and T make UCB's intervals wider than they need to be, while the posterior uses exactly the information available.
- **It generalizes effortlessly.** Any model with a posterior you can sample from (contextual, structured, delayed or batched feedback) gives an algorithm. Guarantees for those settings are harder, and are where the theory is still being written.

Both UCB and Thompson sampling lean on i.i.d. rewards: the confidence interval, the posterior. What if there is no distribution at all?

Adversarial bandits

Exponential weights, with one number per round

Adversarial bandits: the setting

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

adversary fixes losses $\ell_1, \dots, \ell_T \in [-1, 1]^n$ in advance (*oblivious*)

for $t = 1, 2, \dots, T$:

player picks $q_t \in \Delta_n$ and plays $I_t \sim q_t$

player suffers and observes ℓ_{t, I_t} , and

nothing else

$\text{Regret}_T = \max_{i \in [n]} \mathbb{E} \left[\sum_{t=1}^T \ell_{t, I_t} - \ell_{t, i} \right]$,
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 $\ell_1, \dots, \ell_T \in [-1, 1]^n$ in advance (*oblivious*)

for $t = 1, 2, \dots, T$:

player picks $q_t \in \Delta_n$ and plays $I_t \sim q_t$

player suffers and observes ℓ_{t, I_t} , and

nothing else

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

- 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.

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.

The trick: estimate the whole loss vector from one coordinate

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

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

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

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

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

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

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

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

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

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

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

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

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 $\lambda \in \Delta_n$ (e.g. uniform). **Initialize** $w_1 \equiv 1$.

for $t = 1, \dots, T$: $p_t = w_t / \sum_j w_{t,j}$; **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 $-\eta \hat{\ell}_{t,i} \leq 1$ for all i, t ,

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

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 $-\eta \hat{\ell}_{t,i} \leq 1$ for all i, t ,

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

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

$$\begin{aligned} \mathbb{E} \left[\sum_t \ell_{t,I_t} - \ell_{t,i} \right] &= \mathbb{E} \left[\sum_t \langle q_t, \ell_t \rangle - \ell_{t,i} \right] \leq \\ &\leq \end{aligned}$$

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 $-\eta \hat{\ell}_{t,i} \leq 1$ for all i, t ,

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

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

$$\begin{aligned} \mathbb{E} \left[\sum_t \ell_{t,I_t} - \ell_{t,i} \right] &= \mathbb{E} \left[\sum_t \langle q_t, \ell_t \rangle - \ell_{t,i} \right] \leq \gamma T + \mathbb{E} \left[\sum_t (1 - \gamma) \langle p_t, \ell_t \rangle - \ell_{t,i} \right] \\ &\leq 2\gamma T + (1 - \gamma) \mathbb{E} \left[\sum_t \langle p_t, \ell_t \rangle - \ell_{t,i} \right], \end{aligned}$$

using $\gamma \langle \lambda, \ell_t \rangle \leq \gamma$ and $-\ell_{t,i} \leq -(1 - \gamma)\ell_{t,i} + \gamma$. 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 $\mathbb{E}[\sum_t \langle p_t, \hat{\ell}_t \rangle - \hat{\ell}_{t,i}]$: EW's regret *on the estimates*.

Step B: the potential argument

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

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:

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

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:

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

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 $\eta = \sqrt{\log(n)/(3nT)}$:

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

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

Corollary (notes, Cor. 3)

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

$$\max_i \mathbb{E} \left[\sum_{t=1}^T \ell_{t, I_t} - \ell_{t, i} \right] \leq \sqrt{12 n T \log n}.$$

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

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

Corollary (notes, Cor. 3)

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

$$\max_i \mathbb{E} \left[\sum_{t=1}^T \ell_{t, I_t} - \ell_{t, i} \right] \leq \sqrt{12 n T \log n}.$$

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

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

Corollary (notes, Cor. 3)

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

$$\max_i \mathbb{E} \left[\sum_{t=1}^T \ell_{t,I_t} - \ell_{t,i} \right] \leq \sqrt{12 n T \log n}.$$

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

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

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

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

Corollary (notes, Cor. 3)

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

$$\max_i \mathbb{E} \left[\sum_{t=1}^T \ell_{t,I_t} - \ell_{t,i} \right] \leq \sqrt{12 n T \log n}.$$

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

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

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

So $\text{Regret}_T \leq 3\eta n T + \log(n)/\eta = \sqrt{12 n T \log n}$ at $\eta = \sqrt{\log(n)/(3nT)}$. ■

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

Corollary (notes, Cor. 3)

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

$$\max_i \mathbb{E} \left[\sum_{t=1}^T \ell_{t,I_t} - \ell_{t,i} \right] \leq \sqrt{12 n T \log n}.$$

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

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

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

So $\text{Regret}_T \leq 3\eta n T + \log(n)/\eta = \sqrt{12 n T \log n}$ at $\eta = \sqrt{\log(n)/(3nT)}$. ■

The price of bandit feedback, again

Full information $\sqrt{T \log n}$; stochastic $\sqrt{n T \log T}$; adversarial $\sqrt{n T \log n}$. 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. $c = 1/16$).

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)

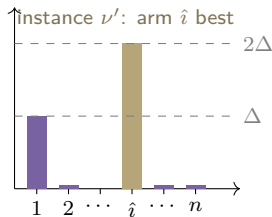
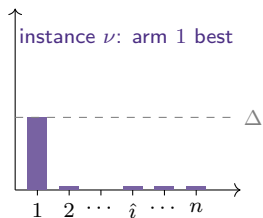
Let $T \geq n \geq 2$. For every policy π there is an instance with Gaussian rewards and means $\theta^* \in [0, 1]^n$ such that

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

for a universal constant c (e.g. $c = 1/16$).

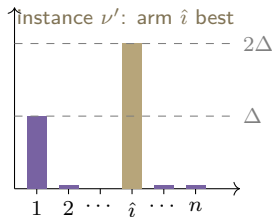
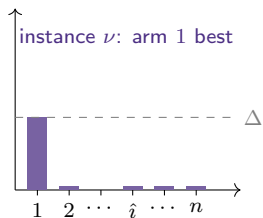
- The quantifiers matter: the hard instance is built *after* seeing the policy. No single instance is hard for every policy.
- The proof is a construction. We will go through the construction and the idea; the two technical facts it rests on we state and do not prove (notes §10.4, Lattimore–Szepesvári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

The construction: a bump where the policy is not looking



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

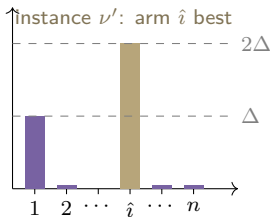
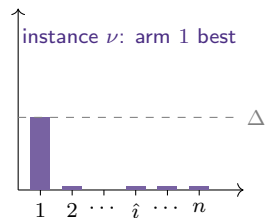
The construction: a bump where the policy is not looking



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

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

The construction: a bump where the policy is not looking

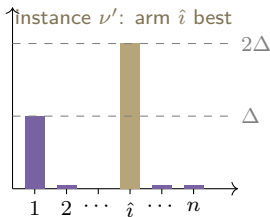
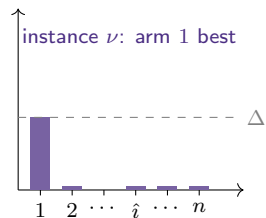


Step 1. Fix the policy π . Instance ν : $\theta^* = (\Delta, 0, \dots, 0)$. Run π on ν (in your head) and record how often it pulls each arm.

Step 2. Some arm $\hat{i} \neq 1$ is pulled least. Pigeonhole:

$$\mathbb{E}_\nu[T_{\hat{i}}] \leq \frac{1}{n-1} \sum_{i \geq 2} \mathbb{E}_\nu[T_i] \leq \frac{T}{n-1}.$$

The construction: a bump where the policy is not looking



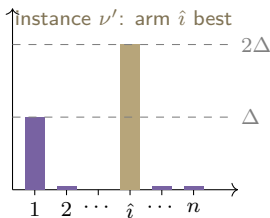
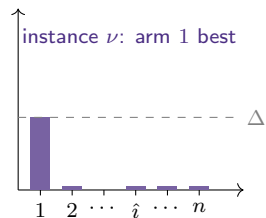
Step 1. Fix the policy π . Instance ν : $\theta^* = (\Delta, 0, \dots, 0)$. Run π on ν (in your head) and record how often it pulls each arm.

Step 2. Some arm $\hat{i} \neq 1$ is pulled least. Pigeonhole:

$$\mathbb{E}_\nu[T_{\hat{i}}] \leq \frac{1}{n-1} \sum_{i \geq 2} \mathbb{E}_\nu[T_i] \leq \frac{T}{n-1}.$$

Step 3. Instance ν' : same as ν , but raise arm $\hat{i}$ to 2Δ . Now $\hat{i}$ 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 : $\theta^* = (\Delta, 0, \dots, 0)$. Run π on ν (in your head) and record how often it pulls each arm.

Step 2. Some arm $\hat{i} \neq 1$ is pulled least. Pigeonhole:

$$\mathbb{E}_{\nu}[T_{\hat{i}}] \leq \frac{1}{n-1} \sum_{i \geq 2} \mathbb{E}_{\nu}[T_i] \leq \frac{T}{n-1}.$$

Step 3. Instance ν' : same as ν , but raise arm $\hat{i}$ to 2Δ . Now $\hat{i}$ 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 :

$$R_T(\pi, \nu) \geq$$

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

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 :

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

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 :

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

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:

$$\begin{aligned} R_T(\pi, \nu) + R_T(\pi, \nu') &\geq \frac{\Delta T}{2} \left(1 - [\mathbb{P}_\nu(E) - \mathbb{P}_{\nu'}(E)] \right) \\ &\geq \end{aligned}$$

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 :

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

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:

$$\begin{aligned} R_T(\pi, \nu) + R_T(\pi, \nu') &\geq \frac{\Delta T}{2} \left(1 - [\mathbb{P}_\nu(E) - \mathbb{P}_{\nu'}(E)] \right) \\ &\geq \frac{\Delta T}{2} \cdot \frac{1}{2} = \frac{T}{8} \sqrt{\frac{n-1}{T}} = \frac{1}{8} \sqrt{(n-1)T}, \end{aligned}$$

provided no event E has probability differing by more than $\frac{1}{2}$ under ν and ν' . Then one of the two regrets is at least $\frac{1}{16} \sqrt{(n-1)T}$.

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 :

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

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:

$$\begin{aligned} R_T(\pi, \nu) + R_T(\pi, \nu') &\geq \frac{\Delta T}{2} \left(1 - [\mathbb{P}_\nu(E) - \mathbb{P}_{\nu'}(E)] \right) \\ &\geq \frac{\Delta T}{2} \cdot \frac{1}{2} = \frac{T}{8} \sqrt{\frac{n-1}{T}} = \frac{1}{8} \sqrt{(n-1)T}, \end{aligned}$$

provided no event E has probability differing by more than $\frac{1}{2}$ under ν and ν' . Then one of the two regrets is at least $\frac{1}{16} \sqrt{(n-1)T}$.

What is left to believe

“ π cannot tell ν from ν' ” means: for every event E about the transcript, $|\mathbb{P}_\nu(E) - \mathbb{P}_{\nu'}(E)| \leq \frac{1}{2}$. Next slide.

Making “cannot tell them apart” precise, in two facts

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

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

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

Making “cannot tell them apart” precise, in two facts

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

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

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

Here only arm $\hat{i}$ differs, by 2Δ , and it is pulled $\leq T/(n-1)$ times:

$$\text{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \leq$$

Making “cannot tell them apart” precise, in two facts

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

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

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

Here only arm $\hat{i}$ differs, by 2Δ , and it is pulled $\leq T/(n-1)$ times:

$$\text{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \leq \frac{T}{n-1} \cdot \frac{(2\Delta)^2}{2} = \frac{2\Delta^2 T}{n-1} = \frac{1}{2} \text{ at } \Delta = \frac{1}{2} \sqrt{(n-1)/T}.$$

Making “cannot tell them apart” precise, in two facts

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

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

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

Here only arm $\hat{i}$ differs, by 2Δ , and it is pulled $\leq T/(n-1)$ times:

$$\text{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \leq \frac{T}{n-1} \cdot \frac{(2\Delta)^2}{2} = \frac{2\Delta^2 T}{n-1} = \frac{1}{2} \text{ at } \Delta = \frac{1}{2} \sqrt{(n-1)/T}.$$

Fact 2: small KL means no event can separate them (Pinsker's inequality)

$$\sup_E |\mathbb{P}_\nu(E) - \mathbb{P}_{\nu'}(E)| \leq \sqrt{\text{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'})/2} \leq \frac{1}{2}.$$

Exactly what the previous slide needed. ■

Making “cannot tell them apart” precise, in two facts

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

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

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

Here only arm $\hat{i}$ differs, by 2Δ , and it is pulled $\leq T/(n-1)$ times:

$$\text{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) \leq \frac{T}{n-1} \cdot \frac{(2\Delta)^2}{2} = \frac{2\Delta^2 T}{n-1} = \frac{1}{2} \text{ at } \Delta = \frac{1}{2} \sqrt{(n-1)/T}.$$

Fact 2: small KL means no event can separate them (Pinsker's inequality)

$$\sup_E |\mathbb{P}_\nu(E) - \mathbb{P}_{\nu'}(E)| \leq \sqrt{\text{KL}(\mathbb{P}_\nu, \mathbb{P}_{\nu'})/2} \leq \frac{1}{2}.$$

Exactly what the previous slide needed. ■

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.

What the lower bound says

- **The hard gaps are $\Delta \approx \sqrt{n/T}$,** exactly the A/B picture with n arms: each suboptimal arm can be examined only T/n times, so a $\sqrt{n/T}$ bump is invisible, yet ignoring it costs $T\sqrt{n/T} = \sqrt{nT}$.
- **Everything today is tight up to logs.** Elimination, UCB and Thompson at $\sqrt{nT \log T}$, EXP3 at $\sqrt{nT \log n}$, and MOSS at $\sqrt{nT}$ exactly. The $\sqrt{n}$ versus $\sqrt{\log n}$ gap between bandits and experts is real.
- **The template recurs.** Two instances that differ only where the algorithm looks least; a KL that adds up over the pulls; a change-of-measure inequality. We will reuse it for linear bandits and active learning.