# A Deeper Dive into Unstructured Bandits

## Principles of Online Decision-Making (CS-303)

Prof. Matthias Grossglauser

Information and Network Dynamics (INDY) lab
School of Computer and Communication Sciences (I&C)
EPFL

# Recap: the UCB algorithm

Notation

- Estimate of mean reward of arm $i$ at time $t$: $\hat{\mu}_i(t)$
- Number of pulls of arm $i$ at time $t$: $T_i(t)$
- Upper confidence bound of arm $i$ at time $t$:

$$UCB_i(t,\delta) = \hat{\mu}_i(t) + \sqrt{2\log(1/\delta)/T_i(t)}$$

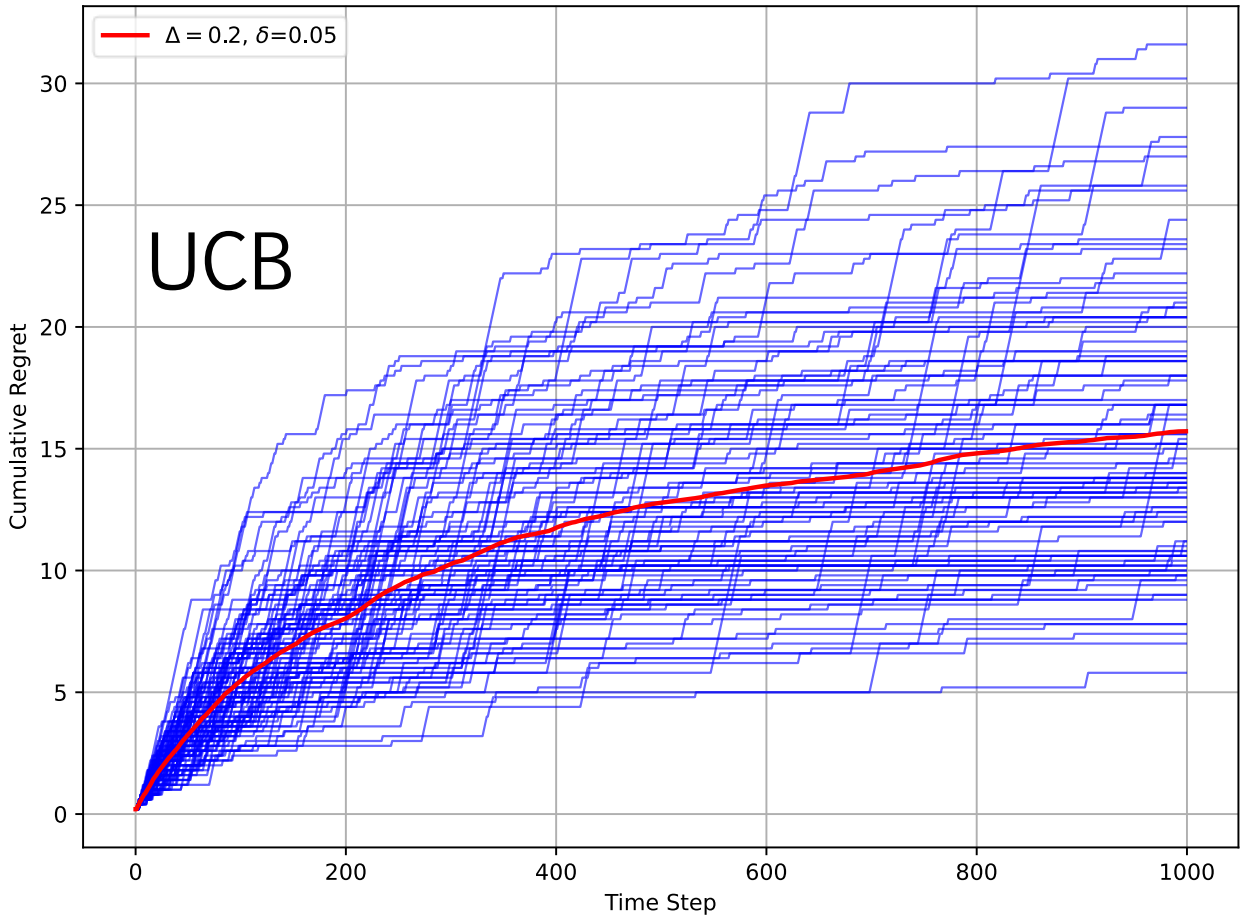
UCB Algorithm

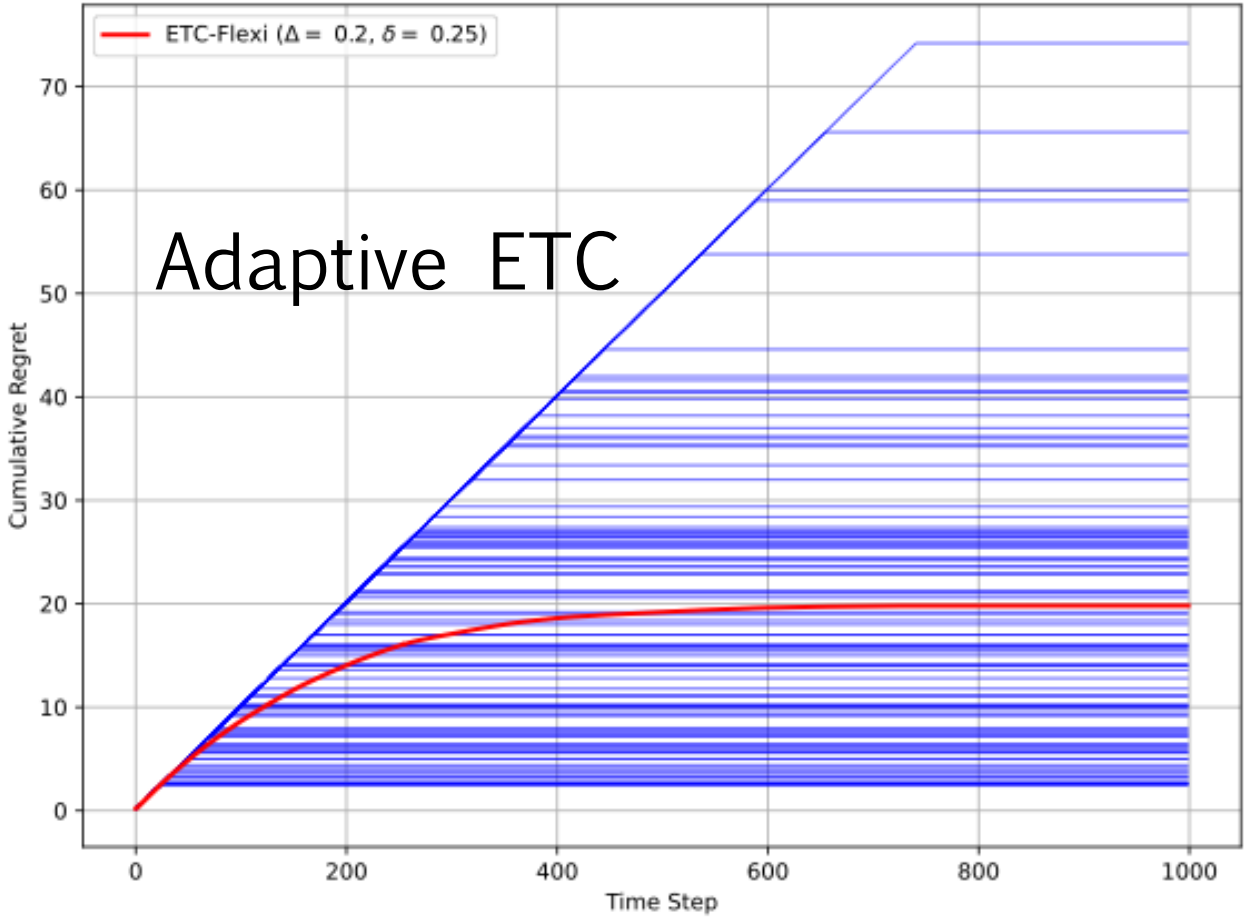
For $t = 1, 2, \ldots, n$ do:

Choose $A_t = \arg\max_i UCB_i(t-1, \delta)$

Observe $X_t$ and update $UCB_i(t, \delta)$

# Recap: exploration-exploitation tradeoff in UCB




In UCB, there is a soft transition from exploration to exploitation, unlike ETC

# Recap: regret bound for UCB

- Instantaneous regret at round $t$: $\mu^* - \mu_{A_t} = \Delta_{A_t}$
- Cumulative regret $R_n = \sum_{t=1}^{n} \mathbb{E}\left[\mu^* - X_{A_t}\right]$
- Regret decomposition: $R_n = \sum_{i \in [k]} \Delta_i \, \mathbb{E}[T_i(n)]$
- Sublinear cumulative regret $\Leftrightarrow$ vanishing instantaneous regret

Theorem:
For the UCB algorithm applied to a stochastic 1-subgaussian bandit with $k$ arms and an error tolerance of $\delta = 1/n^2$,
the regret satisfies

$$R_n \leq 3 \sum_{i=1}^{k} \Delta_i + \sum_{i:\Delta_i > 0} \frac{16 \ln(n)}{\Delta_i}$$

# Recap: regret bound proof strategy

- To bound regret, sufficient to bound $\mathbb{E}T_i(n)$, because $R_n = \sum_i \Delta_i \, \mathbb{E}[T_i(n)]$

Define an event $G_i$:
(1) $\mu_1$ is **always** below its own UCB
(2) $\mu_1$ is above UCB of arm $i$ **after $u_i$ pulls of $i$**

Show that if $G_i$ holds, then $T_i(n) \leq u_i$

Show that if $u_i$ is large enough, then $G_i$ is very likely

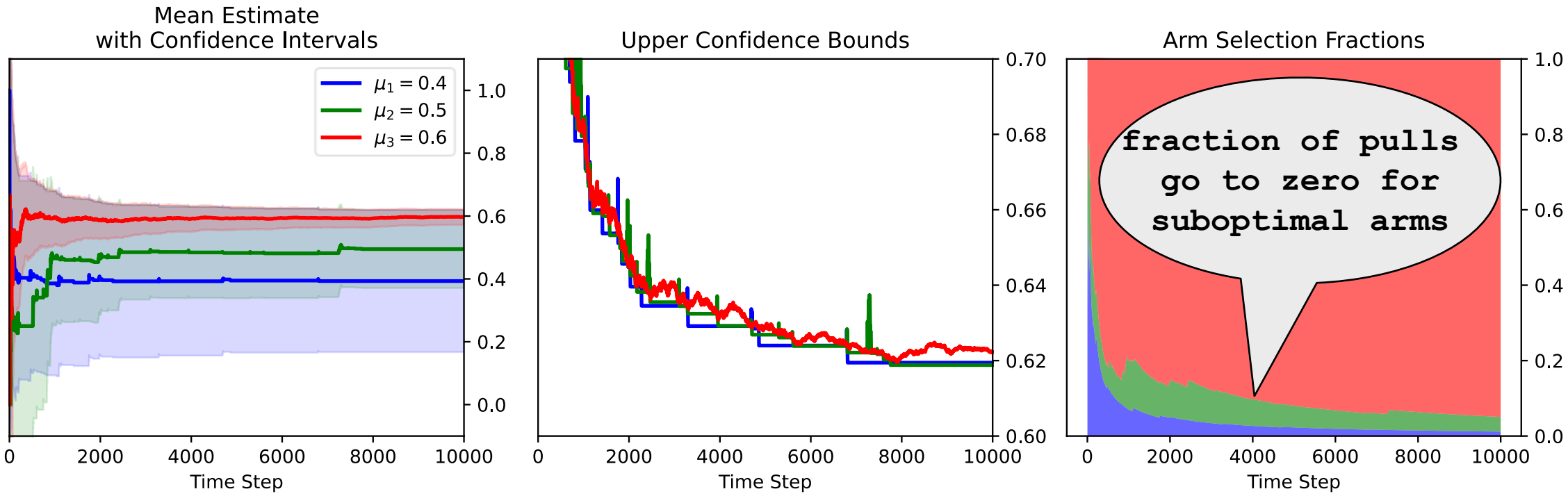
Make a specific choice for $u_i$ that allows to bound $\mathbb{E}T_i(n)$

$$\mathbb{E}T_i(n) \leq 3 + \frac{16 \ln(n)}{\Delta_i^2}$$

$$u_i = \left\lceil \frac{8 \ln(1/\delta)}{\Delta_i^2} \right\rceil$$

# UCB: evolution of confidence intervals

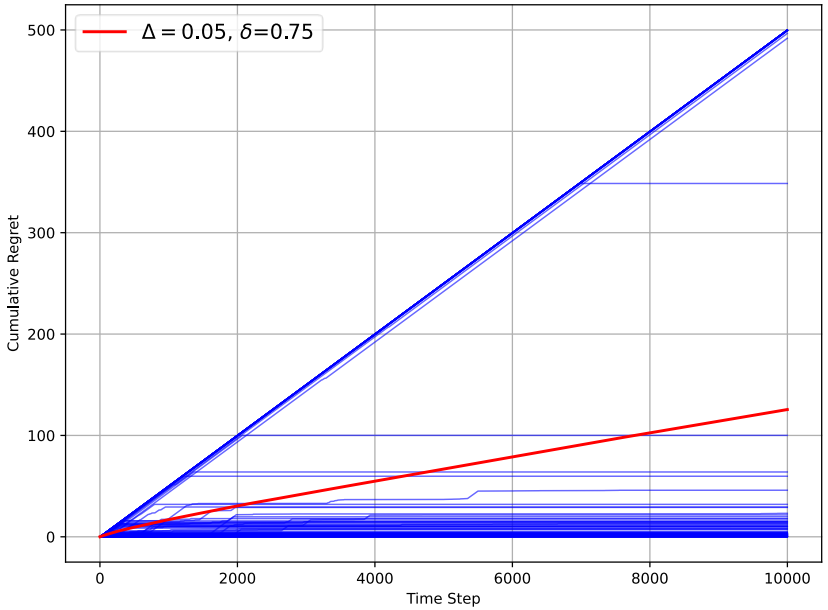
Three-armed Bernoulli bandit with means $[0.4, 0.5, 0.6]$



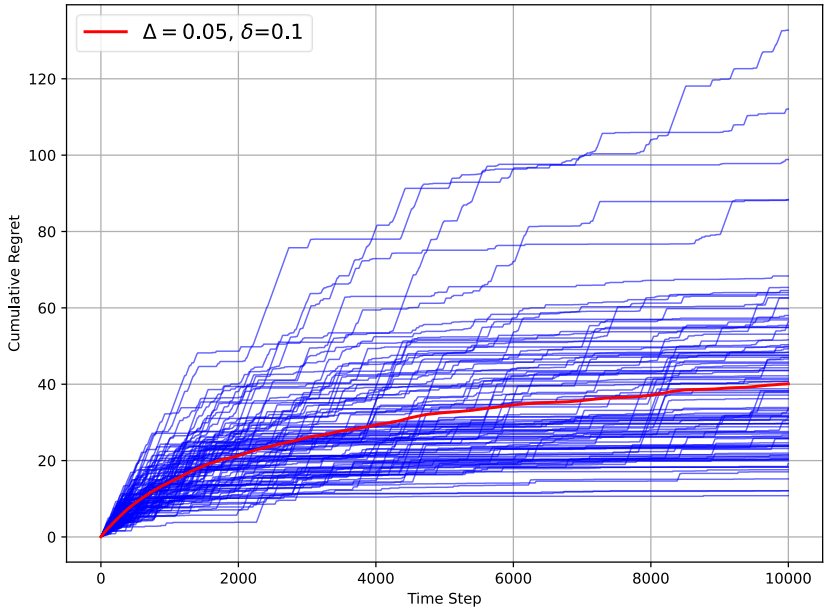

- The upper confidence bounds of the arms are nearly identical throughout
- The worse the arm, the less often it is played

# UCB: the effect of $\delta$

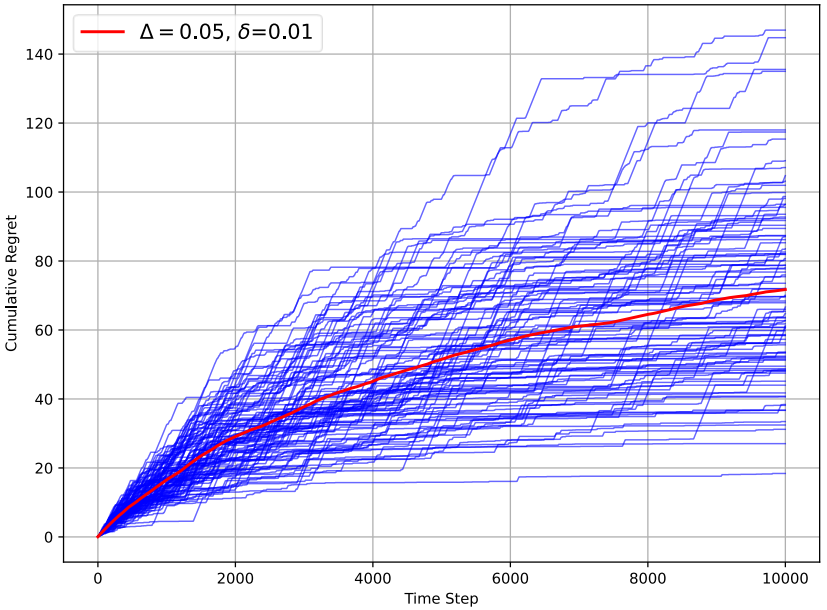
two-armed Bernoulli bandit with means $[0.5 - \Delta/2, 0.5 + \Delta/2]$




under-exploration




correct exploration




over-exploration

- Smaller $\delta$, larger confidence intervals, more exploration
- Too large $\delta$: potential of 'committing' to wrong arm ⇒ large regret
- Too small $\delta$: over-exploration ⇒ large regret

# UCB: alternate regret bound

- We proved the following regret bound:

$$R_n \leq 3 \sum_{i=1}^{k} \Delta_i + \sum_{i:\Delta_i>0} \frac{16 \ln(n)}{\Delta_i}$$

- If $\Delta_i$ arbitrarily small, bound is meaningless
- Can we get a 'problem-independent' bound?

Theorem:
For the UCB algorithm applied to a stochastic 1-subgaussian bandit with $k$ arms and an error tolerance of $\delta = 1/n^2$, the regret satisfies

$$R_n \leq 3 \sum_{i=1}^{k} \Delta_i + 8\sqrt{kn \ln n}$$

we trade $\sum_{i\in[k]} 1/\Delta_i$ for $\sqrt{nk}$

# UCB: alternate regret bound proof

Let's recall the main ingredients of the previous proof:

- Regret decomposition lemma: $R_n = \sum_i \Delta_i \, \mathbb{E}[T_i(n)]$

- Bound derived from concentration inequality: $\mathbb{E}[T_i(n)] \leq 3 + \frac{16 \ln(n)}{\Delta_i^2}$
- Remember: trivial bound: $\mathbb{E}[T_i(n)] \leq n$ (for all arms under all conditions)

- When is the latter bound tighter than the former?

$$\frac{16 \ln(n)}{\Delta_i^2} \geq n \Rightarrow \Delta_i \leq \sqrt{16 \ln(n)/n}$$

Goal is $R_n = O\left(\sqrt{nk \ln(n)}\right)$

- Regret from arm $i$ in this case:

$$\Delta_i \, \mathbb{E}[T_i(n)] = O\left(\sqrt{16 \ln(n)/n}\right) n = O\left(\sqrt{n \ln(n)}\right)$$

# UCB: alternate regret bound proof

Let's recall the main ingredients of the previous proof:

- Regret decomposition lemma: $R_n = \sum_i \Delta_i \, \mathbb{E}[T_i(n)]$
- Bound derived from concentration inequality: $\mathbb{E}[T_i(n)] \leq 3 + \frac{16 \ln(n)}{\Delta_i^2}$
- Remember: a trivial bound: $\mathbb{E}[T_i(n)] \leq n$ (for all arms under all conditions)

Refined strategy:

- Split arms into two sets:
  - Nearly optimal arms: $\Delta_i < \Delta$ (use trivial bound for these arms)
  - Clearly suboptimal arms: $\Delta_i \geq \Delta$ (use concentration bound for these arms)
- Optimize $\Delta$ to get best expression

# UCB: alternate regret bound proof

$$R_n = \sum_i \Delta_i \, \mathbb{E}[T_i(n)] = \sum_{i:\Delta_i < \Delta} \Delta_i \, \mathbb{E}[T_i(n)] + \sum_{i:\Delta_i \geq \Delta} \Delta_i \, \mathbb{E}[T_i(n)]$$

$$\leq n\Delta + \sum_{i:\Delta_i \geq \Delta} \Delta_i \, \mathbb{E}[T_i(n)]$$

$$\leq n\Delta + \sum_{i:\Delta_i \geq \Delta} \Delta_i \left(3 + \frac{16 \ln(n)}{\Delta_i^2}\right)$$

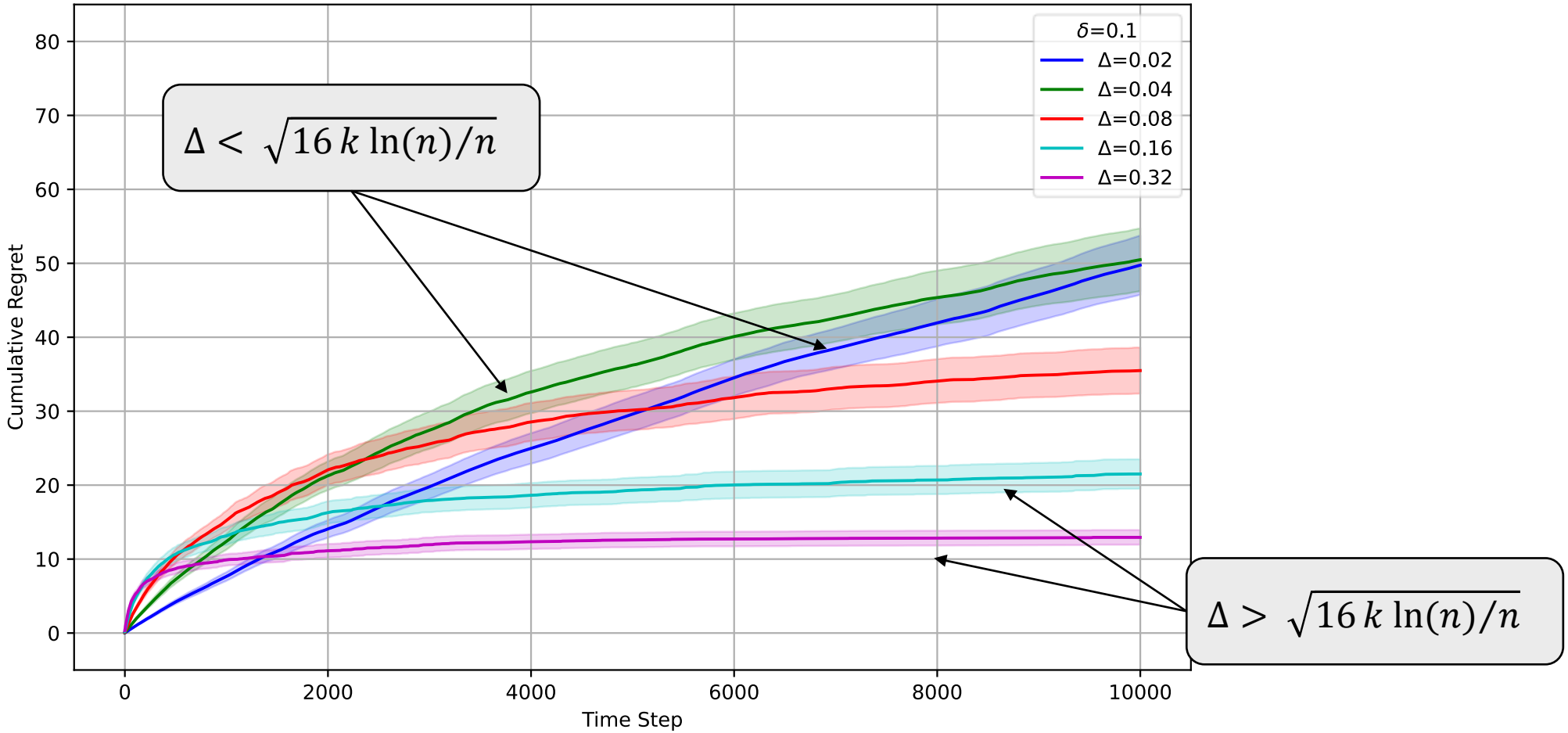
$$\leq \sum_i 3\Delta_i + n\Delta + \frac{16\, k \ln(n)}{\Delta}$$

Merely a proof strategy; no change in algorithm (algorithm does not know suboptimality gaps)

**can use calculus or AM-GM inequality**

Minimised by choosing $\Delta = \sqrt{16\, k \ln(n)/n}$, yielding $R_n \leq \sum_i 3\Delta_i + 8\sqrt{nk \ln(n)}$

# UCB: revisiting regret bound plot




- $O\left(\frac{\ln(1/\delta)}{\Delta^2}\right)$ the number of pulls to separate two arms with gap $\Delta$ w.p. $1-\delta$
- $\Delta$ large $\Rightarrow$ confidence intervals separate earlier $\Rightarrow$ logarithmic regret curves
- $\Delta$ small $\Rightarrow$ later separation $\Rightarrow$ steadily increasing regret curves

# UCB: asymptotic version

- In what we have seen so far, the horizon $n$ is fixed a priori
  - Makes sense in some applications, not in others
- What happens when we do not know the horizon?
- Equivalently, what if we want UCB to run for 'infinite time'?
- Main cause of concern: must eliminate any possibility of persistent error
  - Any error probability, however small, could lead to large regret
- Current formulation:
  - Chance of error $\sim$ chance of concentration event failing $\sim \delta$
  - Cannot choose $\delta = 0$
- In effect, we need $\delta \to 0$ over time

# UCB: asymptotic version

- In effect, need $\delta \to 0$ over time
- Reflected in confidence intervals
  - Originally: $UCB_i(t, \delta) = \hat{\mu}_i(t) + \sqrt{2\log(1/\delta)/T_i(t)}$
  - New formulation: $UCB_i(t) = \hat{\mu}_i(t) + \sqrt{8\log(t)/T_i(t)}$
  - Rest of the algorithm remains identical!
- If an arm is not played for a long time, C.I. slowly grows
  - Ensures an arm cannot be neglected forever
  - Each arm is pulled ever so often
- No 'error' is permanent (mistake in identifying best arm)

# UCB: asymptotic version with regret guarantee

Asymptotic UCB

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

Choose $A_t = \arg\max_i \left( \hat{\mu}_i(t-1) + \sqrt{\frac{8 \ln(t)}{T_i(t-1)}} \right)$

Observe $X_t$ and update $UCB_i(t)$ for all $i$

Theorem:

For any 1-subgaussian bandit, the regret of the asymptotic UCB algorithm satisfies

$$R_n \leq \sum_{i:\Delta_i > 0} \left( 4\Delta_i + \frac{32 \ln(n)}{\Delta_i} \right)$$

$$\limsup_{n \to \infty} \frac{R_n}{\ln(n)} \leq \sum_{i:\Delta_i > 0} \frac{32}{\Delta_i}$$

# Asymptotic UCB: regret bound sketch

- Like in fixed horizon version:
  - The key is to prove $\mathbb{E}[T_i(n)] \leq \left(32 \ln(n)/\Delta_i^2\right) + 4$ for suboptimal arms
- We will follow broadly the same steps:
  - Define good events such that:
    - Under good event ("normal circumstances"), number of pulls of suboptimal arm is bounded
    - Probability of bad event is small (concentration inequality)
- Differences from fixed horizon case:
  - Good event now depends on time $t$
  - Careful use of union bound

# Defining good events

## Fixed Horizon Setting

- $G = G_1 \cap G_i$, where
  - $G_1 = \left\{\mu_1 < \min_{t\in[n]} \hat{\mu}_1(t) + \sqrt{\frac{2}{T_1(t)} \ln\left(\frac{1}{\delta}\right)}\right\}$
  - $G_i = \left\{\hat{\mu}_{iu_i} + \sqrt{\frac{2}{u_i} \ln\left(\frac{1}{\delta}\right)} < \mu_1\right\}$

- Conditional on $G$
  $\Rightarrow T_i(n) \leq \left\lceil \frac{8 \ln(1/\delta)}{\Delta_i^2} \right\rceil \triangleq u_i$

## Asymptotic Setting

- $G_t = G_{1,t} \cap G_{i,t}$, where
  - $G_{1,t} = \left\{\mu_1 < \hat{\mu}_1(t) + \sqrt{\frac{8}{T_1(t)} \ln(t)}\right\}$
  - $G_{i,t} = \left\{\mu_i > \hat{\mu}_i(t) - \sqrt{\frac{8}{T_i(t)} \ln(t)}\right\}$

- Conditional on $G_t$
  $\Rightarrow T_i(t) \leq \left\lceil \frac{32 \ln(n)}{\Delta_i^2} \right\rceil$
  **OR**
  arm $i$ not pulled in round $t$

**proof on next slide**

# Defining good events

- $G_t = G_{1,t} \cap G_{i,t}$, where
  - $G_{1,t} = \left\{ \mu_1 < \hat{\mu}_1(t) + \sqrt{\frac{8}{T_1(t)} \ln(t)} \right\}$
  - $G_{i,t} = \left\{ \mu_i > \hat{\mu}_i(t) - \sqrt{\frac{8}{T_i(t)} \ln(t)} \right\}$

$UCB$ of arm 1 above true mean at time $t$

$LCB$ of arm $i$ below true mean at time $t$

- Need to show:

$$G_t \cap \{A_t = i\} \Rightarrow T_i(t) \leq \left\lceil \frac{32 \ln(n)}{\Delta_i^2} \right\rceil$$

- Equivalently:

$$G_t \cap \left\{ T_i(t) > \left\lceil \frac{32 \ln(n)}{\Delta_i^2} \right\rceil \right\} \Rightarrow \{A_t \neq i\}$$

$(A \Rightarrow B) \Rightarrow (B^c \Rightarrow A^c)$

$(C \cap A \Rightarrow B) \Rightarrow (C \cap B^c \Rightarrow A^c)$

# Defining good events

- Assume both $G_t$ and $T_i(t) > \left\lceil \frac{32 \ln(n)}{\Delta_i^2} \right\rceil$ occur

$$\hat{\mu}_i(t) - \sqrt{\frac{8 \ln(t)}{T_i(t)}} + \Delta_i < \hat{\mu}_1(t) + \sqrt{\frac{8 \ln(t)}{T_1(t)}}$$

$$T_i(t) > \left\lceil \frac{32 \ln(n)}{\Delta_i^2} \right\rceil \Rightarrow \Delta_i > 2\sqrt{\frac{8 \ln(t)}{T_i(t)}}$$

Combining: $\underbrace{\hat{\mu}_i(t) + \sqrt{\frac{8 \ln(t)}{T_i(t)}}}_{UCB_i(t)} < \underbrace{\hat{\mu}_1(t) + \sqrt{\frac{8 \ln(t)}{T_1(t)}}}_{UCB_1(t)}$

$\Rightarrow$ arm $i$ **not pulled** in round $t$

# Bounding $\mathbb{E}[T_i(n)]$

## Fixed Horizon Setting

- $\mathbb{E}[T_i(n)]$

$\leq \mathbb{E}[\mathbb{I}\{G\}T_i(n)] + \mathbb{E}[\mathbb{I}\{G^c\}T_i(n)]$

$\leq u_i + \mathbb{P}(G^c)n$

Too loose for asymptotic version!

- $\mathbb{P}(G_i^c) \leq n\delta + \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)$
- Put together, gives desired result

## Asymptotic Setting

- $\mathbb{E}[\mathbb{I}\{A_t = i\}] = \mathbb{E}[\mathbb{I}\{G_t\}\mathbb{I}\{A_t = i\}] + \mathbb{E}[\mathbb{I}\{G_t^c\}\mathbb{I}\{A_t = i\}]$

$\leq \mathbb{E}[\mathbb{I}\{T_i(t) \leq u_i\}\mathbb{I}\{A_t = i\}] + \mathbb{E}[\mathbb{I}\{G_t^c\}]$

- $\mathbb{E}[T_i(n)] = \sum_{t\in[n]} \mathbb{E}[\mathbb{I}\{A_t = i\}]$

$\leq \sum_{t\in[n]} \mathbb{E}[\mathbb{I}\{T_i(t) \leq u_i\}\mathbb{I}\{A_t = i\}] + \mathbb{E}[\mathbb{I}\{G_t^c\}]$

$\leq u_i + \sum_{t\in[n]} \mathbb{P}(G_t^c)$

$\leq u_i + \sum_{t=1}^{\infty} \mathbb{P}(G_t^c)$

How many times can one increment a counter while it's below 17?

$\leq \frac{32\ln(n)}{\Delta_i^2} + 1 + \sum_{t=1}^{\infty} \frac{2}{t^3} \leq \frac{32\ln(n)}{\Delta_i^2} + 4$

$\sum_{t=1}^{\infty} \frac{1}{t^3} \leq 1 + \int_1^{\infty} \frac{1}{t^3} dt = 1.5$

# Bounding $\mathbb{E}[T_i(n)]$

- Need to show $\mathbb{P}\left(G_{1,t}^c\right) \leq t^{-3}$
- Same bound for $\mathbb{P}\left(G_{i,t}^c\right)$
- Together, we get $\mathbb{P}(G_t^c) \leq \mathbb{P}\left(G_{1,t}^c\right) + \mathbb{P}\left(G_{i,t}^c\right) \leq 2t^{-3}$
- $G_{1,t}^c = \left\{\mu_1 \geq \hat{\mu}_1(t) + \sqrt{\frac{8\ln(t)}{T_1(t)}}\right\} \subseteq \bigcup_{s=1}^{t} \left\{\mu_1 \geq \hat{\mu}_{1,s} + \sqrt{\frac{8\ln(t)}{s}}\right\}$

take union over all possible values of $T_1(t)$

- $\mathbb{P}\left(G_{1,t}^c\right) \leq \sum_{s=1}^{t} \mathbb{P}\left(\mu_1 \geq \hat{\mu}_{1,s} + \sqrt{\frac{8\ln(t)}{s}}\right) = \sum_{s=1}^{t} \mathbb{P}\left(\mu_1 \geq \hat{\mu}_{1,s} + \sqrt{\frac{2\ln(1/t^{-4})}{s}}\right) \leq \sum_{s=1}^{t} t^{-4} = t^{-3}$

Chernoff bound

# Asymptotically Optimal UCB

- We have proven the result

$$\limsup_{n\to\infty} \frac{R_n}{\ln(n)} \leq \sum_{i:\Delta_i>0} \frac{32}{\Delta_i}$$

- Algorithm 6 and Theorem 8.1 in L&S gives asymptotically **optimal** result:

$$\limsup_{n\to\infty} \frac{R_n}{\ln(n)} \leq \sum_{i:\Delta_i>0} \frac{2}{\Delta_i}$$

- Differences:
  - Exploration bonus is $\sqrt{\frac{2\ln(f(t))}{T_i(t-1)}}$; $f(t) = (1 + t\,(\ln t)^2)$ instead of $t^4$
  - Good events measure UCB crossing the threshold $\mu_1 - \varepsilon$ rather than $\mu_1$
  - Homework: read Chapter 8 of L&S

# UCB algorithm: a long history

- 1985: Lai and Robbins: A clean formulation of the multi-armed bandit problem.
  - Proved fundamental logarithmic lower bound
  - Also introduced the idea of optimism and using confidence bounds

- 1985–1995: UCB algorithm proposed, asymptotic regret bounds

- 2002: Auer et al. proposed UCB1 algorithm and analysed its regret
  - This is what we saw in class today

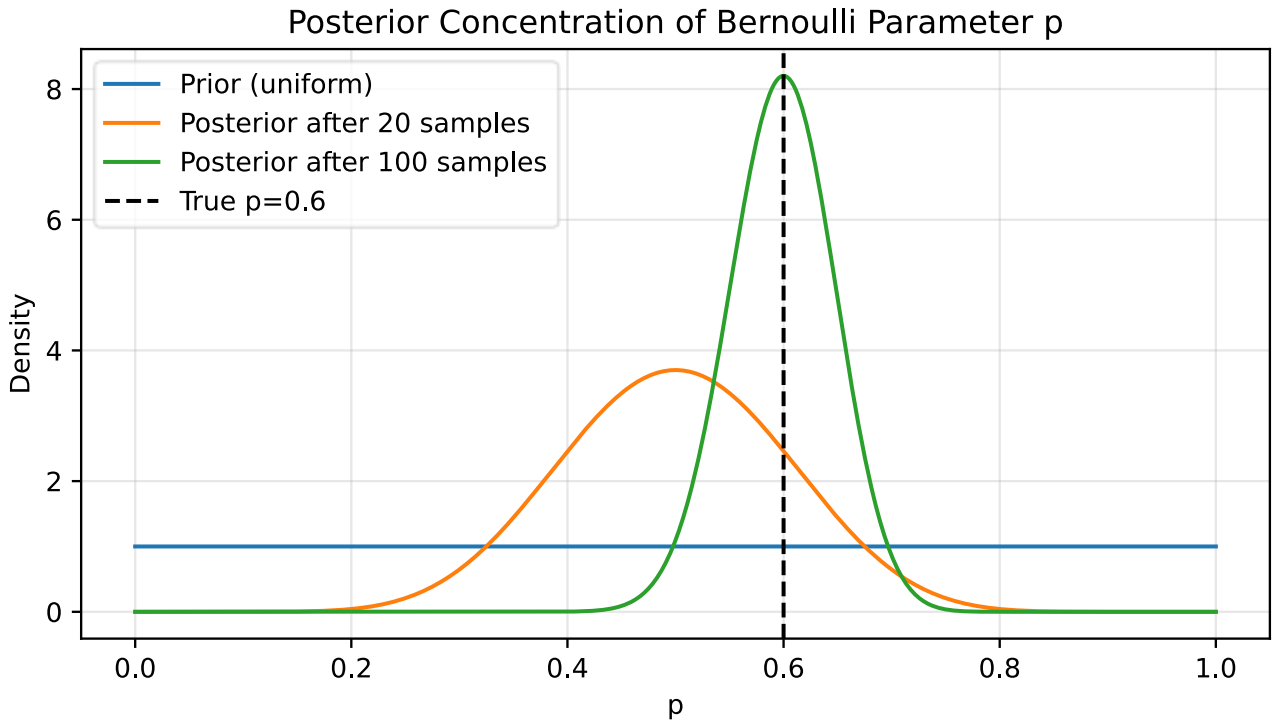
- 2010–2017: Asymptotically Optimal and Minimax Optimal algorithms: MOSS, KL_UCB, etc

- Coming up: Thompson Sampling, an algorithm from 1933!

# Prelude: Bayesian statistics

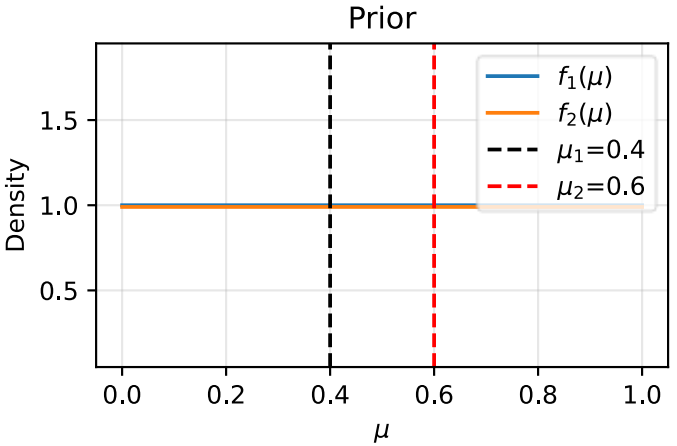
- In Bayesian statistics, we maintain a 'belief distribution' over unknown quantities
  - Belief distribution measures all we know about the unknown
  - Without any data, belief distribution has large variance
  - With more data, belief distribution concentrates around true parameter

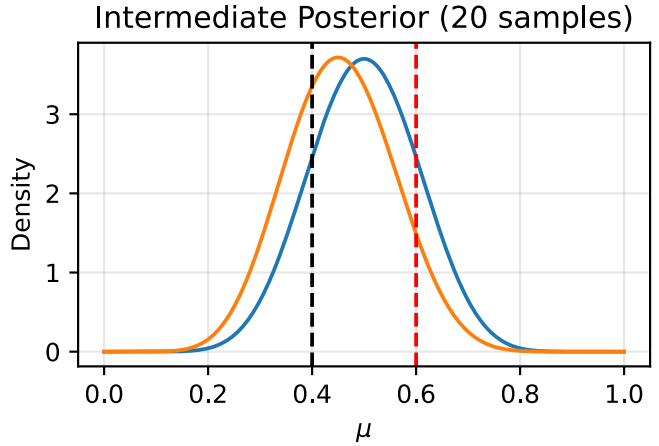
- Example:

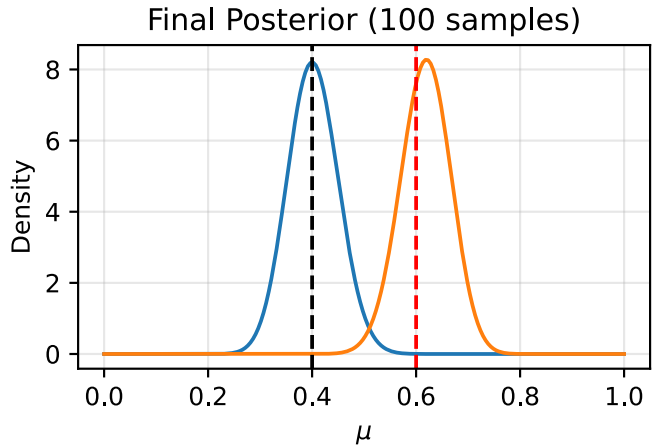


# Bayesian view of bandits

- In multi-armed bandits with $\{0,1\}$ reward, the unknown variables are mean rewards of each arm: $\mu_1, \mu_2, \ldots \mu_k$
- Maintain a belief $f_{i,t}(\mu)$ for all arms $i$, at all times $t$
- Initially, we know nothing about $\mu_i$
  - Any value in $[0,1]$ is equally likely
  - The **prior belief** $f_{i,0}(\mu)$ for each arm $i$ is the uniform distribution on $[0,1]$
- Over time, we pull arms and observe rewards,
  - The belief $f_{i,t}(\cdot)$ concentrates near $\mu_i$



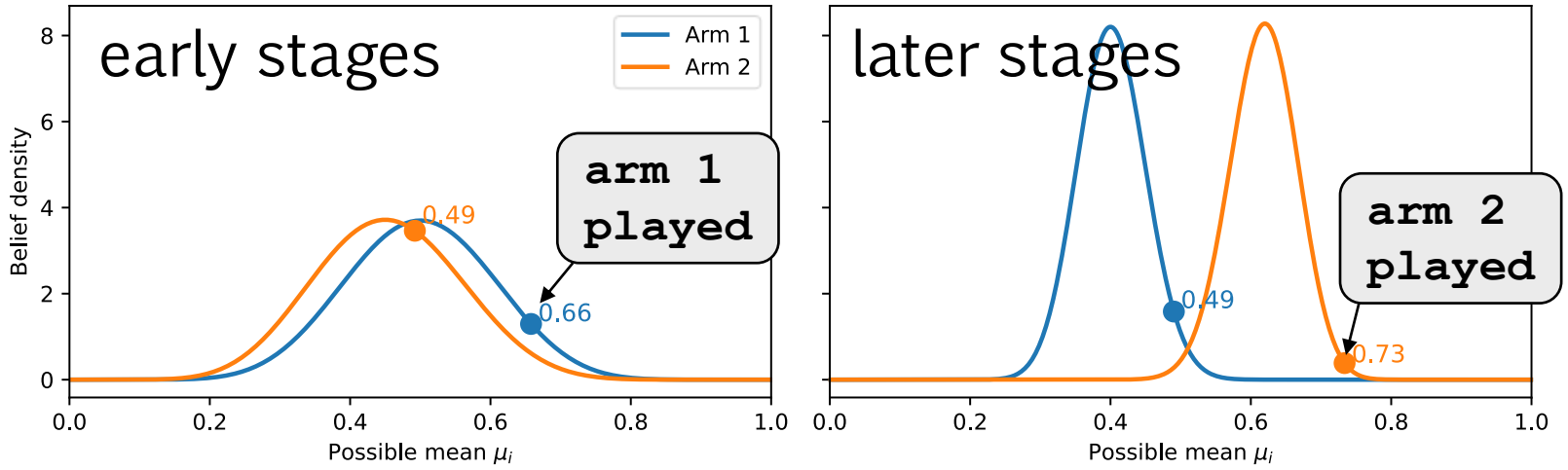



# Thompson Sampling: an introduction

- Thompson sampling is a randomised algorithm for multi-armed bandits

- Algorithm maintains 'belief distribution' of possible arm values at all times
  - Contrast: UCB maintains a point estimate and confidence interval of arm values

- At each step, the algorithm
  - Sample values $\tilde{\mu}_1, \tilde{\mu}_2, \ldots \tilde{\mu}_k$ according to belief distributions
  - Play the arm with the highest value (optimism)
  - Observe data and update belief



# Thompson sampling: algorithm and intuition

Thompson Sampling

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

  For $i = 1, 2, \ldots, k$ do:

    sample $\tilde{\mu}_i(t) \sim f_{i,t}(\cdot)$

  Choose $A_t = \arg\max_i \left(\tilde{\mu}_i(t)\right)$

  Observe $X_t$, update $f_{A_t,t}(\cdot)$

- Exploration happens through the randomness in the sampling
- Initially, beliefs are weaker. larger variance → more exploration
- Over time, beliefs concentrate, samples close to true mean → exploitation
- Smooth transition from exploration to exploitation

# Thompson sampling: evolution of posterior

Arm 1 has smaller reward

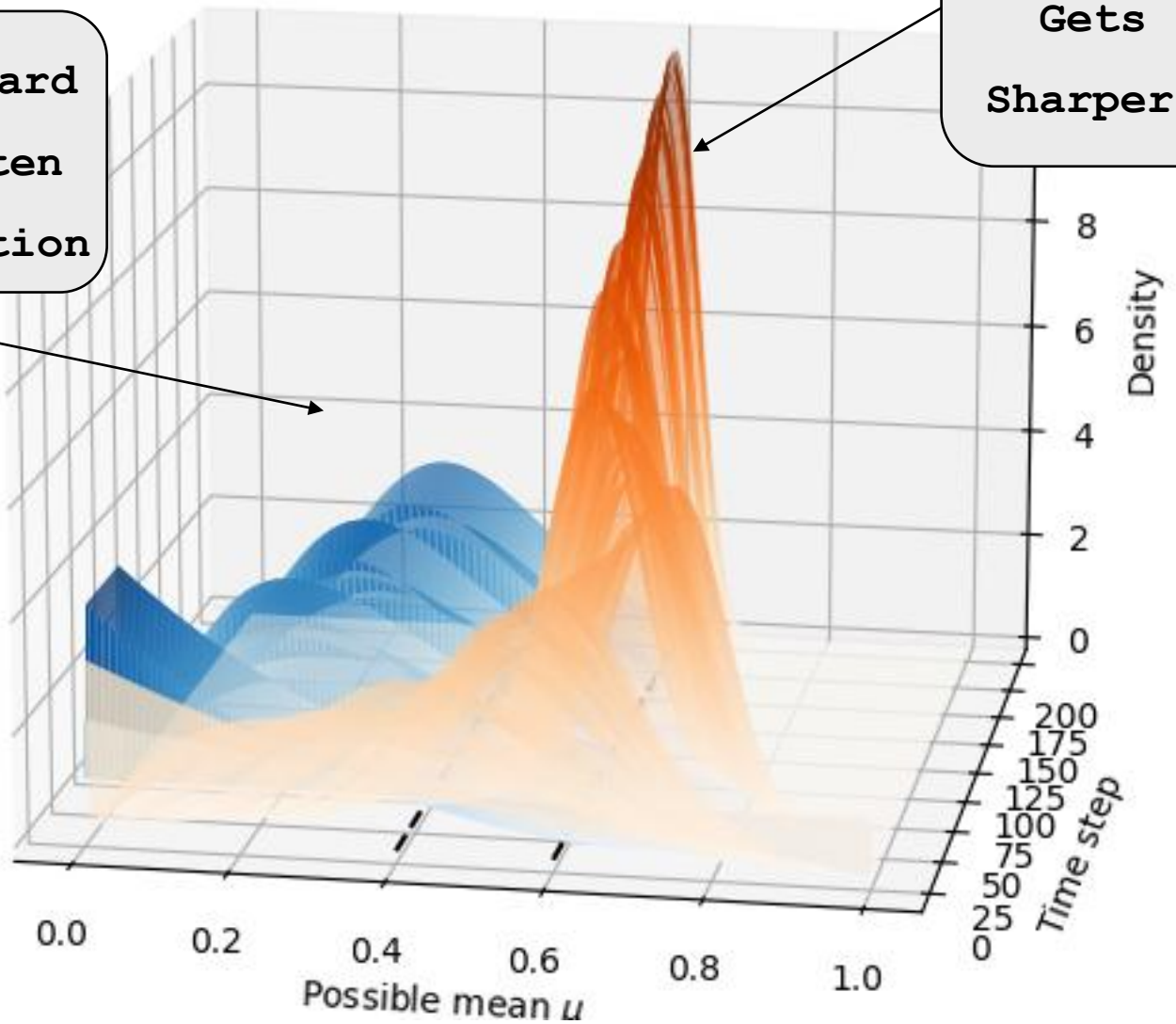
Gets sampled less often

Wider belief distribution

Arm 2 has higher reward

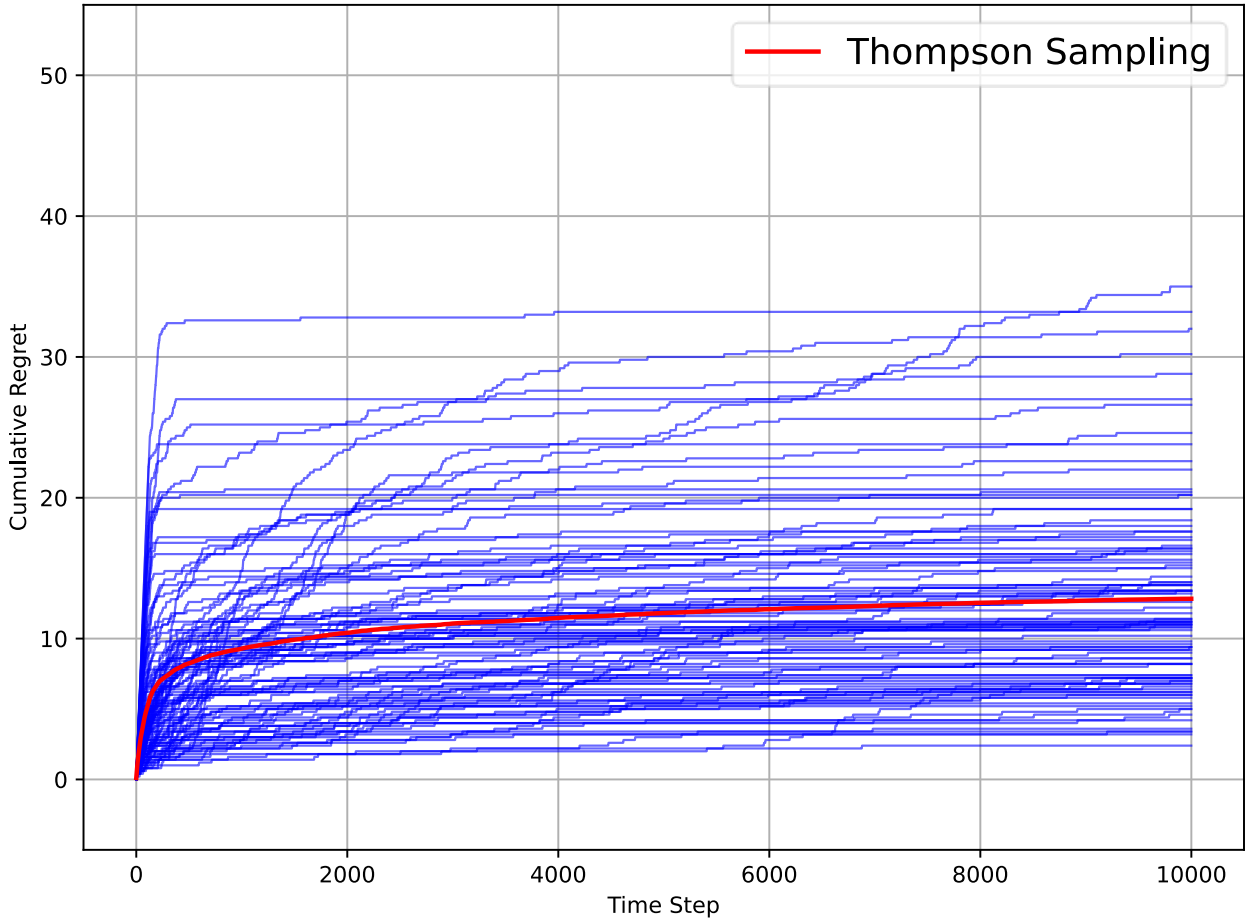
Gets sampled more often

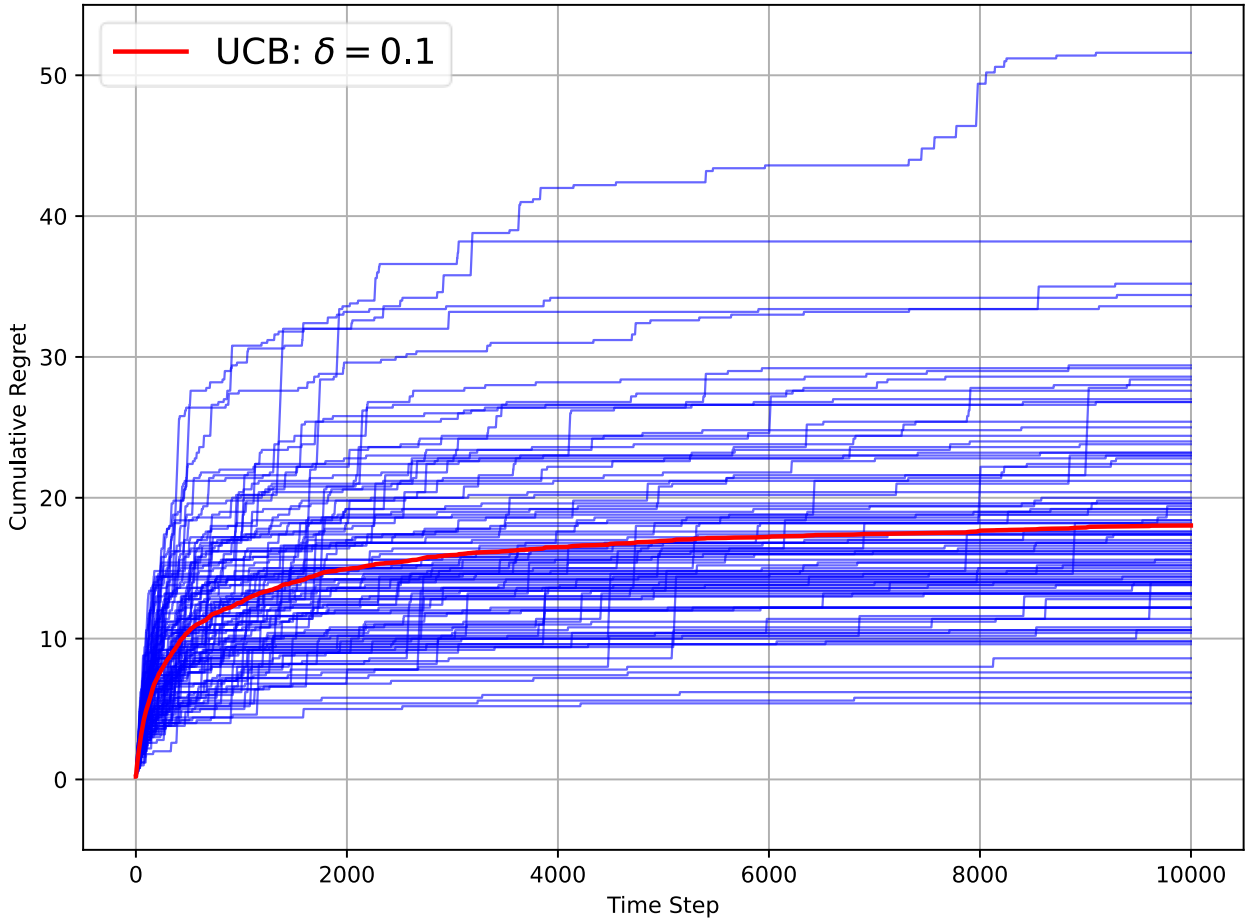
Sharper belief distribution



# Thompson sampling versus UCB

2-armed Bernoulli bandit with means $[0.4, 0.6]$




- Regret curves of Thompson Sampling similar to UCB
  - smooth exploration-exploitation tradeoff
- In practice, Thompson sampling can be better than UCB

# Calculating Posteriors

In the **Bernoulli bandit** setting (rewards 0 or 1),

we use a convenient conjugate prior – the **Beta distribution**:

$$\mu_i \sim Beta(\alpha_i, \beta_i)$$

- $\alpha_i$: count of past successes,
- $\beta_i$: "pseudo-count" of past failures.

**Beta distribution pdf:** $\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

After observing a reward from $r_t \in \{0,1\}$ from arm $i$:

- $\alpha_i \leftarrow \alpha_i + r_t,\ \beta_i \leftarrow \beta_i + (1 - r_t)$,
- the **posterior** remains a Beta distribution
- updating beliefs is just incrementing these two counts

- More data → posterior becomes **narrower** (more confident).
- Successes shift the density **rightward;** failures shift it **leftward**.

# Some practical points

- Thompson sampling has similar theoretical regret guarantees as UCB
  - In practice, can sometimes achieve lower regret than UCB
  - Usually competitive in performance
  - Benefits seen more often in linear bandits (next lecture)

- Computing posterior can be hard without conjugate priors!
  - many techniques to sample from un-normalised distribution, e.g. MCMC
  - other techniques to approximate posterior distribution

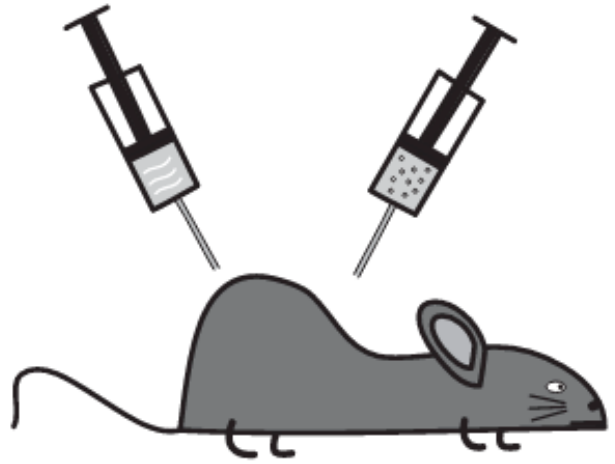
- No hyperparameter to adjust exploration-exploitation trade-off
  - In practice, can speed up or slow down the update of the posterior
  - Or artificially inflate or reduce the variance of the belief distribution

- Thompson sampling, like UCB is a method to minimise regret in bandits
  - But what if we are only interested in *identifying the best arm*?

# Best-arm identification: motivation

Motivation

- Suppose a company wants to test efficacy of drugs on mice
- Few different drug variants → arms
- We do not care about the outcome of drug on mice (no regret)
- Only goal is to identify the best drug with high probability at end of trial!
- How should the drugs be tried?



Source: L&S, Chapter 33

# Best-arm identification: formulation

Basic formulation:

- After $n$ rounds of playing a multi-arm bandit, identify the best arm with high probability
- Return arm $i$ such to minimise $\mathbb{P}(\mu^* - \mu_i > 0)$

Slightly weaker requirement:

- return arm $i$ such that $\mathbb{P}(\mu^* - \mu_i > \epsilon) \leq \delta$
- PAC: probably approximately correct

Yet another alternative:

- minimise simple regret at time $n+1$: $\mathbb{E}[\mu^* - \mu_i]$
- Different from cumulative regret minimization

# Best-arm identification: intuition

Why is this different from regret minimization?

- regret minimizing algorithm play suboptimal arms $O(\log(n))$ times
- confidence intervals are chosen such that error probability $\delta = O(n^{-2})$
- no exploration-exploitation tradeoff: pure exploration

Consider uniform exploration

- Each arm is pulled $\lfloor n/k \rfloor$ times, return best arm after round $n$
- Using Chernoff bound and union bound:

$$\mathbb{P}(\mu^* - \mu_i > 0) \leq \sum_{i:\Delta_i>0} \mathbb{P}(\hat{\mu}_i \geq \hat{\mu}_1) \leq \sum_{i:\Delta_i>0} \exp\left(-\frac{\lfloor n/k \rfloor \Delta_i^2}{4}\right)$$

- Can we do better?

# Sequential halving algorithm

Sequential Halving Algorithm for Best Arm Identification

Input: number of arms $k$, time horizon $n$

Set number of phases $\lceil \log_2(k) \rceil$, active set $\mathcal{A}_1 = [k]$.

For $\ell = 1, 2, \ldots L$ do:

Let $T_\ell = \left\lfloor \frac{n}{L|\mathcal{A}_\ell|} \right\rfloor$

Play each arm in $\mathcal{A}_\ell$ exactly $T_\ell$ times

For each arm $i \in \mathcal{A}_\ell$, compute $\hat{\mu}_i^\ell$: empirical mean of last $T_\ell$ samples

$\mathcal{A}_{\ell+1} \leftarrow$ top $\lceil |\mathcal{A}_\ell|/2 \rceil$ arms in $\mathcal{A}_\ell$

Return $\mathcal{A}_{L+1}$

- Eliminates the worst half of the arms in each round
- More efficient use of experiment budget

# Performance of sequential halving algorithm

Theorem:

The sequential halving algorithm, applied to a stochastic 1-subgaussian bandit with $k$ arms with mean vector $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_k$ for $n$ rounds yields

$$\mathbb{P}(\mu^* - \mu_i > 0) \leq 3\log_2(k)\exp\left(-\frac{n}{16H_2(\mu)\log_2(k)}\right); \qquad H_2(\mu) = \max_{i:\Delta_i>0} \frac{i}{\Delta_i^2}$$

- Case 1: all suboptimal arms have same suboptimality gap
  - $\mu_1 = 1, \mu_2 = \mu_3 = \cdots = \mu_k = 1 - \Delta \;\Rightarrow\; H_2(\mu) = k/\Delta^2$
  - error rate for sequential halving $\sim (n\Delta^2)/(k\log k)$
  - uniform exploration is marginally better $\sim (n\Delta^2)/k$
- Case 2: one arm is slightly worse, all other far worse
  - $\mu_1 = 1, \mu_2 = 1 - \Delta,\ \mu_3 = \cdots = \mu_k = 0 \;\Rightarrow\; H_2(\mu) = 2/\Delta^2$
  - error rate for sequential halving $\sim (n\Delta^2)/\log k$
  - much better than uniform exploration, especially for large $k$

# Summary

- Alternate regret bound: $O(\sqrt{n})$
  - Do not use concentration inequality for small gap arms
- Asymptotic UCB (Chapter 8)
  - Confidence intervals grow with time
  - Over time, chance of error should go to zero
  - Similar regret result and proof as last time
- Thompson Sampling (Chapter 36)
  - Bayesian perspective
  - Randomised algorithm
  - Can be better than UCB in practice
- Best arm identification (Chapter 33)
  - A pure exploration problem
  - Pull suboptimal arms more often than in regret minimization

# Reading assignment

- L&S:
  - Rest of Chapter 7
  - Chapter 8 (but our result is slightly looser and therefore simpler, you can also just stick to the slides)
  - Chapter 36: just read the introduction
  - Chapter 33: Sections 33.1 and 33.3