The Upper Confidence Bound Algorithm
(or: a Lecture on Optimism)

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: multi-armed bandits

- Problem Setup
  - k arms, each arm generates random rewards
  - Arm i has reward distribution P_i, mean μ_i
  - Best arm has mean μ^*
  - Reward distributions are unknown

- Algorithm
  - At time t, take action A_t
  - A_t chosen from set of arms: $1, 2, \ldots, k$
  - Observe reward $X_t \sim P_{A_t}$
  - Continue for $t = 1, 2, \ldots, n$
  - Sequential decision-making process

- Goal: achieve sublinear regret
  - Eventually pull the best arm every time

Recap: regret in bandits

- Regret definition
  - Instantaneous regret at round t: $\mu^* - \mu_{A_t} = \Delta_{A_t}$
  - Δ_{A_t} measures suboptimality of arm A_t
  - bad arm ⇔ large regret
  - good arm ⇔ small regret
  - best arm ⇔ zero regret
  - Cumulative regret $R_n = \sum_{t=1}^{n} \mathbb{E}\left[\mu^* - X_{A_t}\right]$

- Regret decomposition: $R_n = \sum_{a \in \mathcal{A}} \Delta_a \, \mathbb{E}[T_a(n)]$

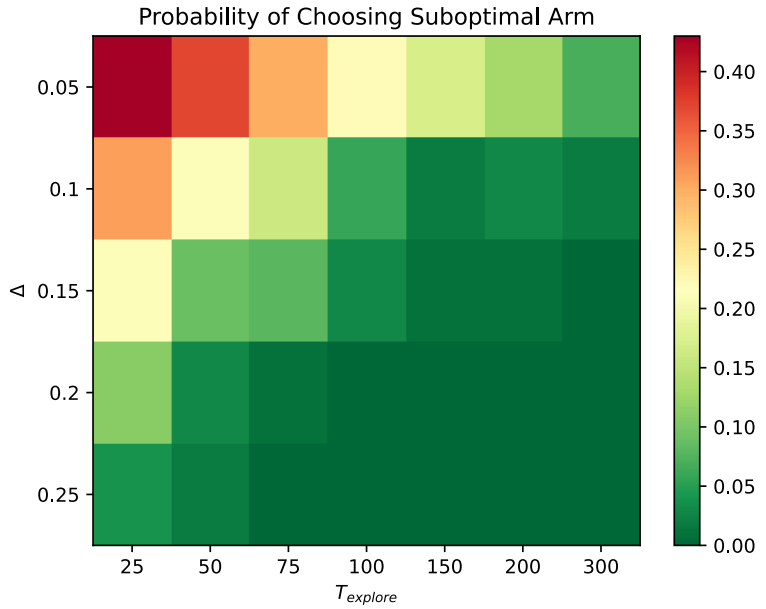
- Sublinear cumulative regret ⇔ vanishing instantaneous regret
  - To minimize regret, eliminate bad arms quickly
  - Eventually identify the best arm(s) and pull that
  - Due to randomness, cannot identify good/bad arms with certainty!

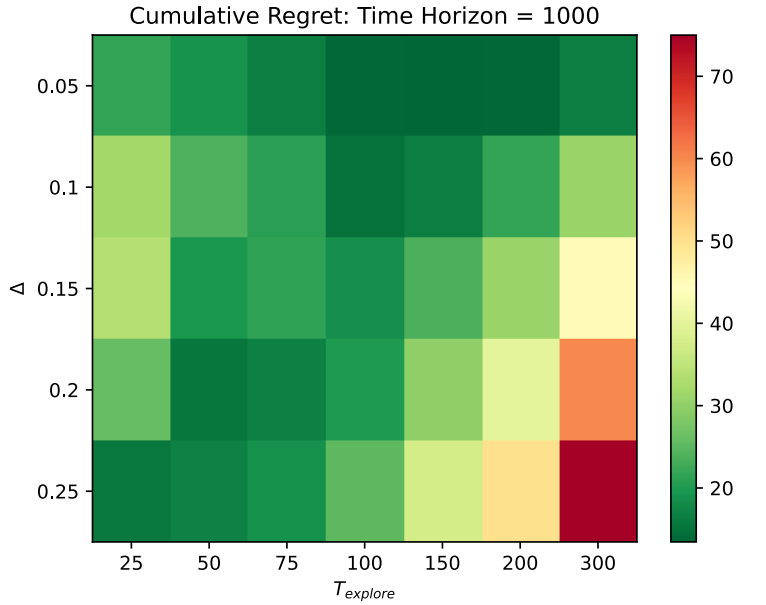
Recap: Explore-Then-Commit (ETC) algorithm

- Algorithm overview:
  - Explore until some fixed time km
  - Exploration phase gives an estimate $\hat{\mu}_i$ of μ_i for all i
  - If $\hat{\mu}_i \approx \mu_i$, picking $i^* = \text{argmax}_i \hat{\mu}_i$ identifies best arm
  - Commit (keep pulling) i^* for rounds $km + 1$ to n

- Intuition
  - If best arm is correctly identified, no regret for commit phase $\Rightarrow R_n \propto \tau$
  - If suboptimal arm is chosen, constant positive regret at all rounds $\Rightarrow R_n \propto n$

- Tradeoff
  - Very small $m \Rightarrow |\hat{\mu}_i - \mu_i|$ large $\Rightarrow$ likely to choose suboptimal arm $\Rightarrow$ large regret
  - Very large $m \Rightarrow$ Excessive exploration $\Rightarrow$ large regret
  - Need to know Δ to choose good m





Adaptive ETC

- Key idea: think of m as a stopping time (instead of fixed)
  - m can depend on data
  - Signals when to stop exploring and start exploiting
  - Ideally, stop when one is sufficiently confident about the optimal arm

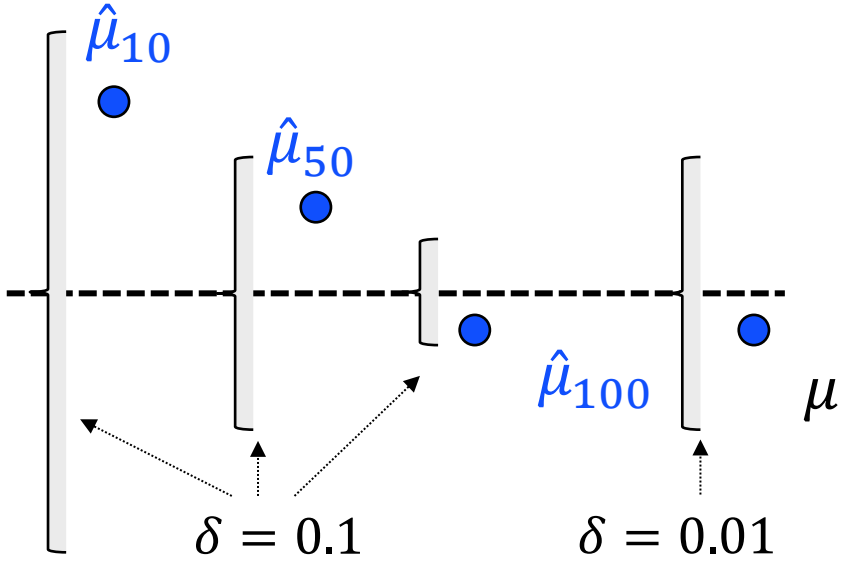
- With fixed m
  - It is possible that one is not sure about optimal arm (**under-exploration**)
  - Or, one was very sure about optimal arm much earlier (**over-exploration**)

- Can decide when to stop by looking at the data
  - Data-dependent exploration rounds m
  - Keep pulling both arms equally and calculating $\hat{\mu}_i$, until one is (reasonably) certain that one arm is better than the other

- How to quantify certainty?
  - Confidence intervals!

Confidence intervals

- Quantifying (un)certainty in mean estimates
  - $X_1, X_2, \ldots, X_n$ are i.i.d. samples of r.v. with unknown mean μ
  - $\hat{\mu}_n = (\sum_i X_i)/n$: estimate of μ
  - Typically, $\hat{\mu}_n$ will be close to μ, but not exactly μ
  - We can say: $|\hat{\mu}_n - \mu| \leq \epsilon$ with probability $1 - \delta$
  - Equivalently, $\mu \in [\hat{\mu}_n - \epsilon, \hat{\mu}_n + \epsilon]$ with high probability



- Confidence Interval (ϵ) v/s Confidence Level (δ)
  - Suppose $X_1, X_2, \ldots, X_n$ are subgaussian random variables
  - Given ϵ, can calculate δ using concentration inequality
  - Alternatively, fix δ and calculate ϵ
  - For fixed δ, ϵ reduces as n increases
  - For fixed n, ϵ reduces as δ increases
  - How are n, ϵ, δ related to subgaussian parameter σ?

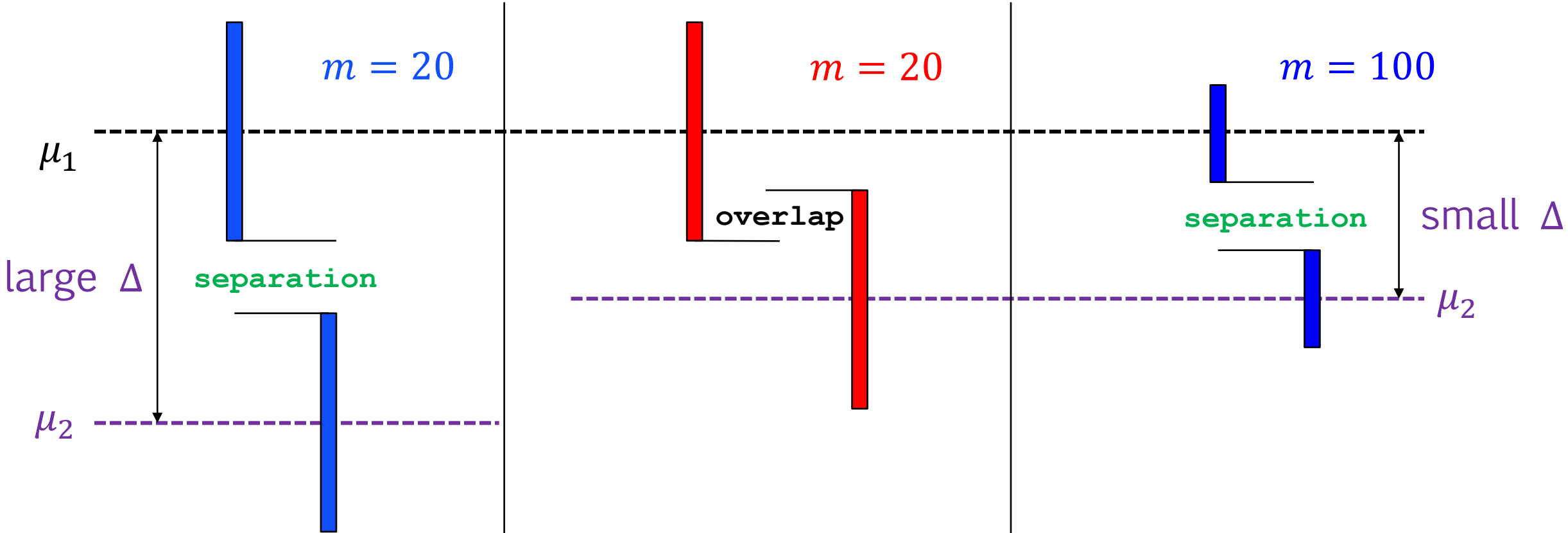
Subgaussian Concentration Inequality

$$\mathbb{P}[|\hat{\mu}_n - \mu| \geq \epsilon] \leq 2\exp\left(-n\frac{\epsilon^2}{2\sigma^2}\right)$$

δ

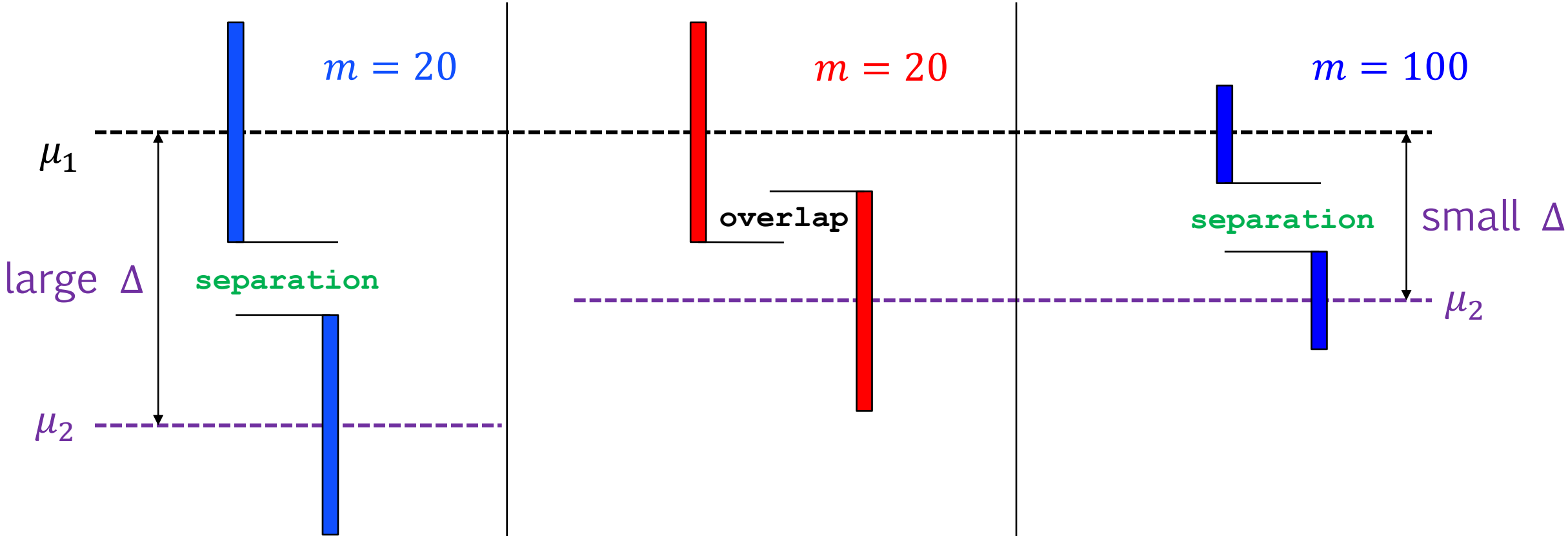
Using confidence intervals for adaptive ETC

- How can we confidently decide when one arm is better than the other?
  - Pretend that each arm's reward lies in the interval with certainty
  - If confidence intervals overlap, best arm is undecided
  - If confidence intervals are disjoint, best arm is clear!
  - Strategy: keep exploring until confidence intervals separate



Adaptive ETC

- Data-dependent m
  - With sufficient exploration, confidence intervals will eventually separate
  - When $\mu_1 - \mu_2$ is large, confidence intervals separate early $\Rightarrow m$ is small
  - When $\mu_1 - \mu_2$ is small, confidence intervals separate after long time $\Rightarrow m$ is large
  - Exploration time is data-dependent, not fixed a priori



Adaptive ETC

- Data-dependent m
  - With sufficient exploration, confidence intervals will eventually separate
  - When $\mu_1 - \mu_2$ is large, confidence intervals separate early $\Rightarrow m$ is small
  - When $\mu_1 - \mu_2$ is small, confidence intervals separate after long time $\Rightarrow m$ is large
  - Exploration time is data-dependent, not fixed a priori

- Notation for Confidence Intervals:
  - UCB: Upper Confidence Bound → $\hat{\mu} + \epsilon$
  - LCB: Lower Confidence Bound → $\hat{\mu} - \epsilon$

- Typically, will compute ϵ as a function of n, δ (and σ)
  - Start from $\mathbb{P}[|\hat{\mu}_n - \mu| \geq \epsilon] \leq 2\exp\left(-n\frac{\epsilon^2}{2\sigma^2}\right)$
  - $UCB(n,\delta) = \hat{\mu}_n + \sigma\sqrt{\frac{2\ \log(2/\delta)}{n}}$, $LCB(n,\delta) = \hat{\mu}_n - \sigma\sqrt{\frac{2\ \log(2/\delta)}{n}}$

ETC with Confidence Intervals: pseudo code

Initialize: $T_a(0) = 0$, $\hat{\mu}_a(0) = 0$ for all arms $a \in \mathcal{A}$. `committed_arm = None`
Hyperparameters: $\delta \in (0,1)$, minimum pulls m_0

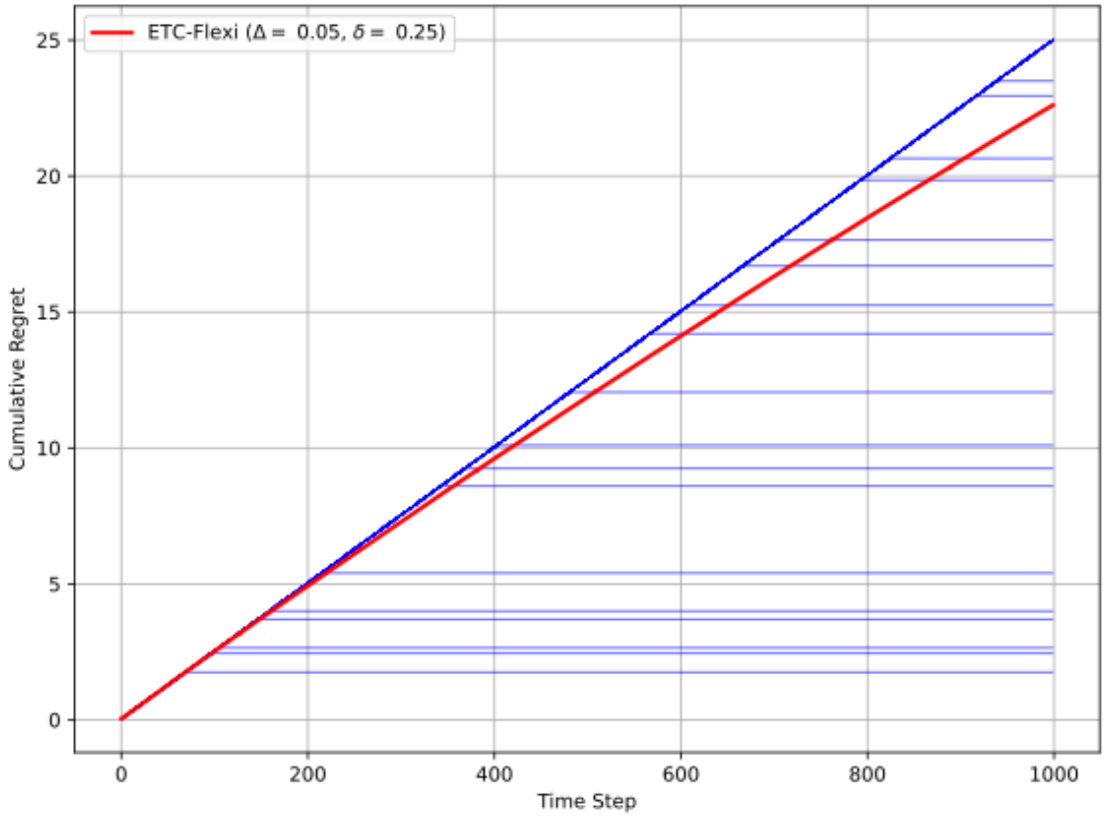
For $t = 1, 2, \ldots, n$ do:

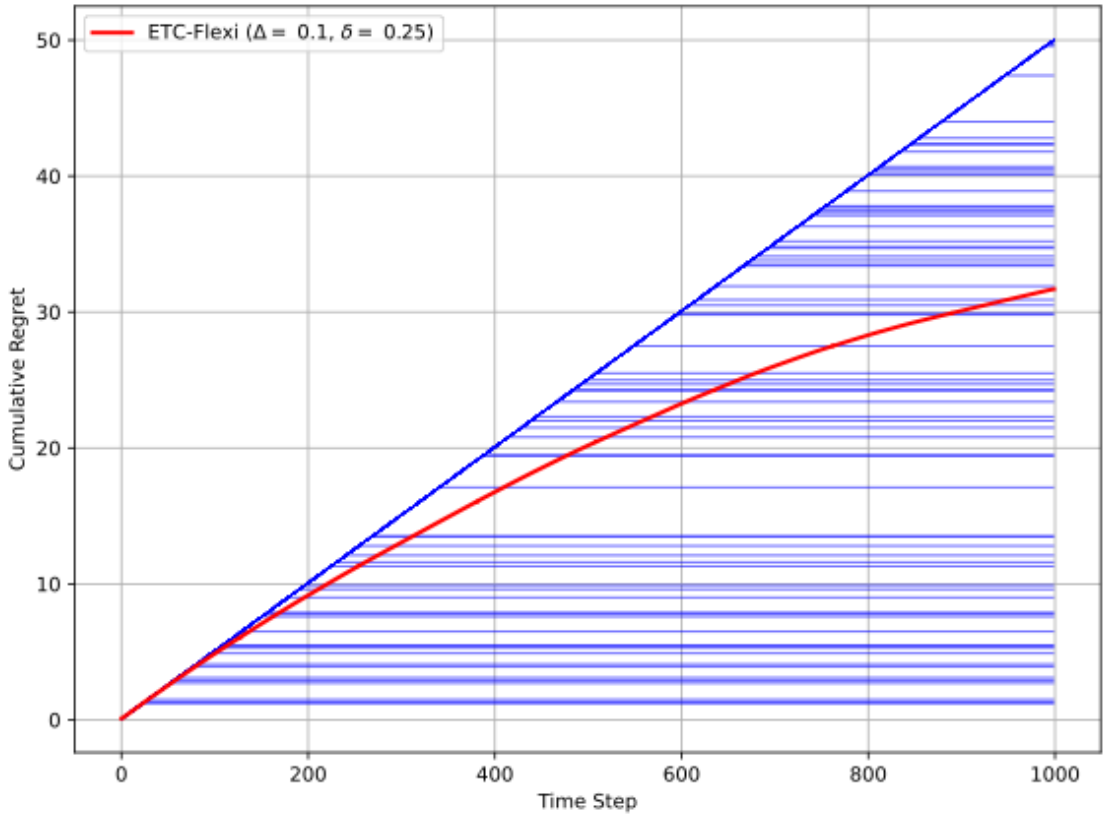
If `committed_arm` $\neq$ `None`, $A_t =$ `committed_arm`
Else: $A_t = (t \bmod k) + 1$

Check for commitment if $t \geq m_0 k$:
Compute $\hat{\mu}_a(t)$ for all arms $a \in \mathcal{A}$; $a^*(t) = \arg\max_a \hat{\mu}_a(t)$
Compute $UCB_a(t, \delta)$ and $LCB_a(t, \delta)$ for all arms $a \in \mathcal{A}$

If $LCB_{a^*}(t, \delta) > UCB_a(t, \delta)$ for all $a \neq a^*$,
`committed_arm` $= a^*$

Adaptive ETC: regret curves

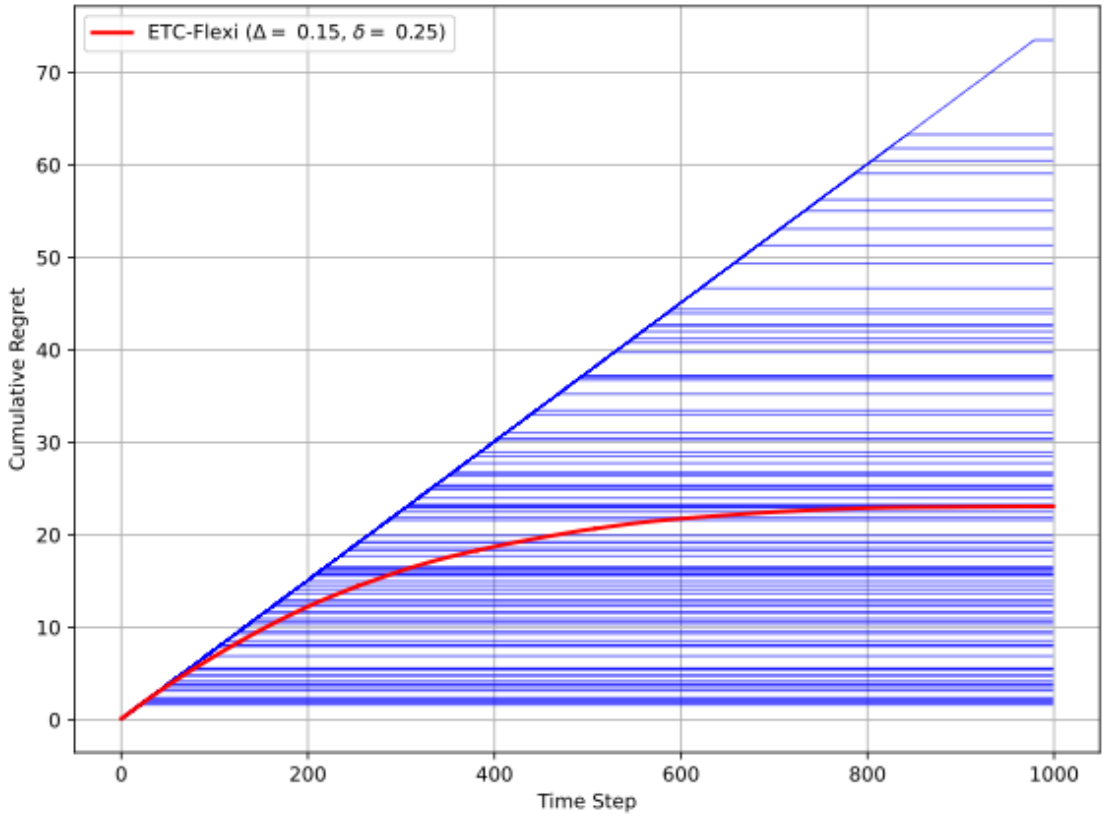


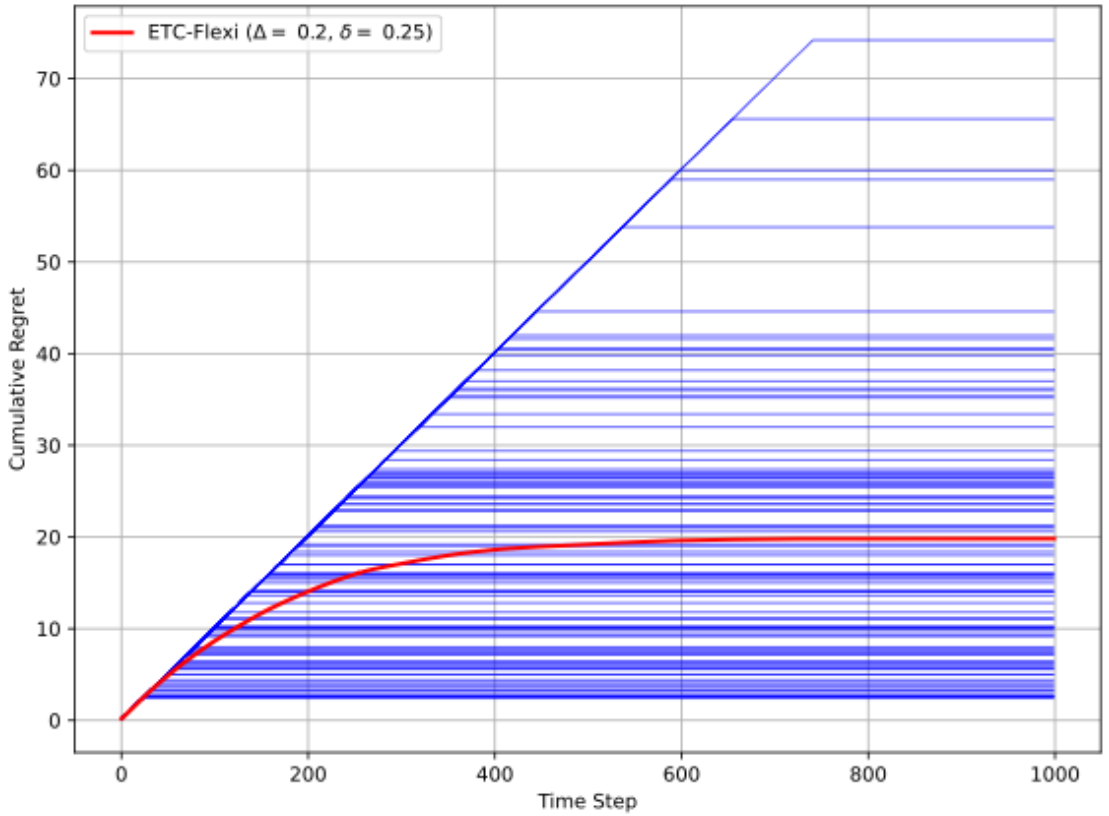


Salient Features:

- Exploration period is random
- Expected cumulative regret is sublinear

Adaptive ETC: regret curves

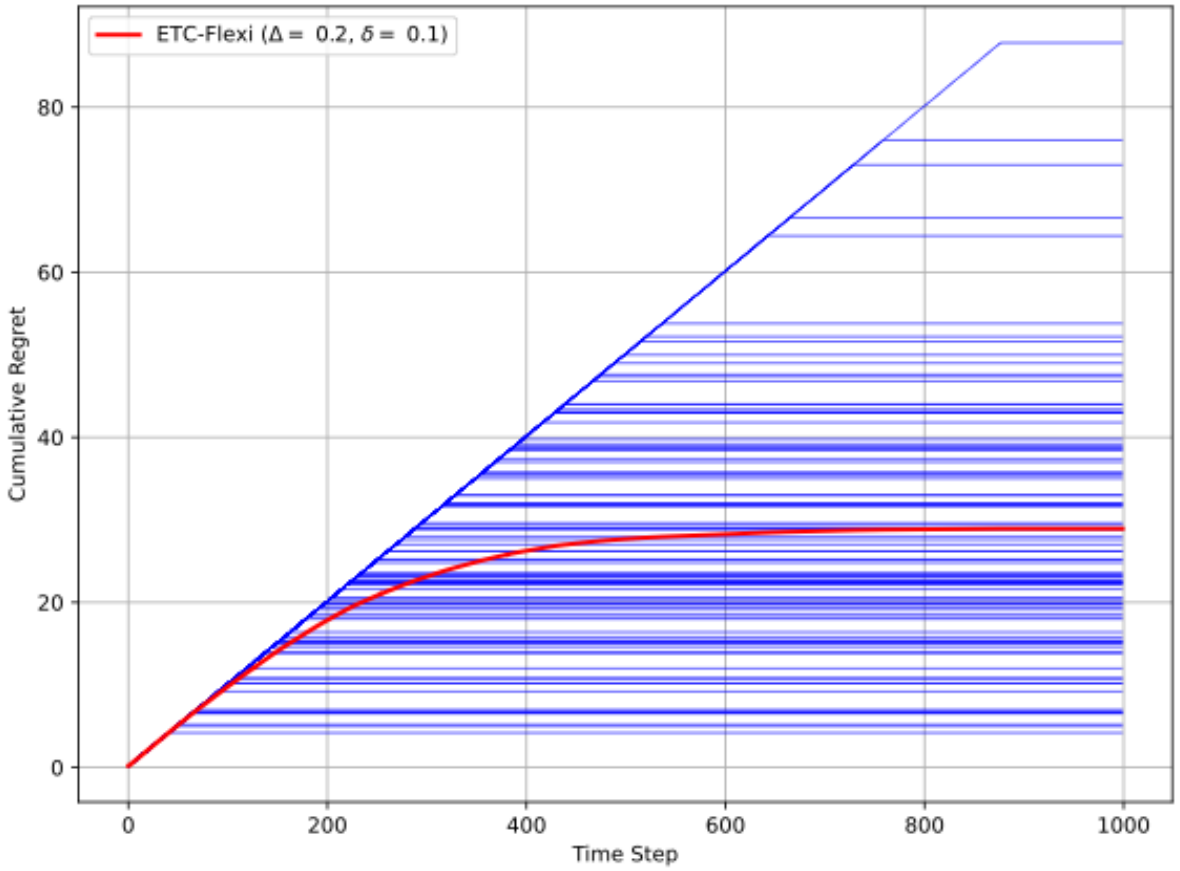


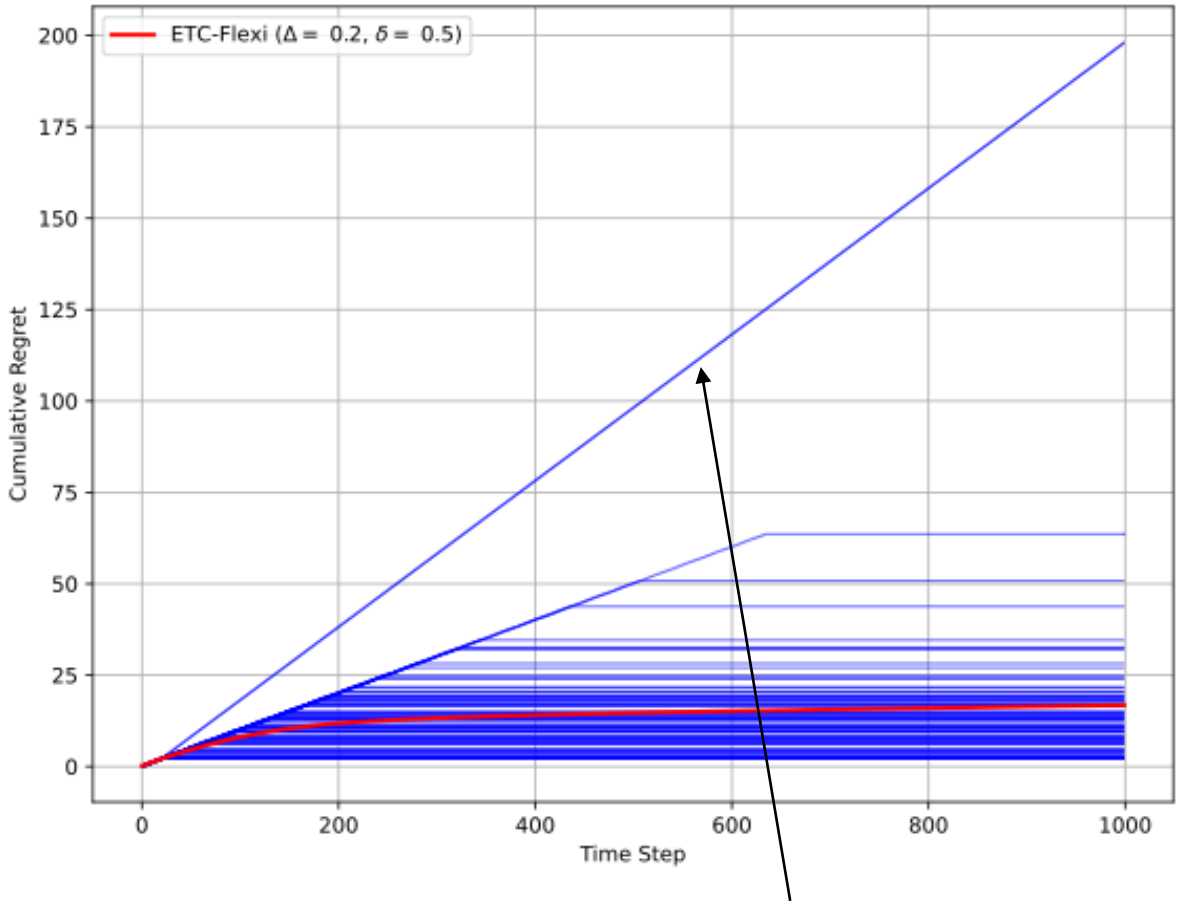


Salient Features:

- Larger $\Delta \Rightarrow$ algorithm commits earlier
- Like for fixed ETC, regret is highest for moderate Δ (0.1)

Adaptive ETC: regret curves





commits to wrong arm

Importance of hyperparameter δ:

- Very small $\delta \Rightarrow$ too conservative (over-exploration)
- Very large $\delta \Rightarrow$ too optimistic (chance of mistaken commitment)
- Rule of thumb: larget the horizon n, smaller the δ (mistakes more costly)

ETC for $k > 2$ arms

- How can we extend ETC principle to multi-armed bandits?
  - Need to pull each arm a few times, to estimate arm's reward
  - Exploration of all arms is necessary

- Similar to two-armed bandits...
  - With more exploration, confidence intervals for each arm keeps shrinking
  - Eventually, confidence intervals will become disjoint
  - Best arm becomes clear

- Different interpretations of ETC
  - View 1: keep exploring until best arm is identified
    - Commit to arm i when its C.I. dominates C.I. of all other arms
  - View 2: keep eliminating arms that are clearly not the best
    - If C.I. of any arm i dominates C.I. of arm j, stop pulling arm j
  - Equivalent with two arms, different for multi-armed case!

From ETC to sequential elimination

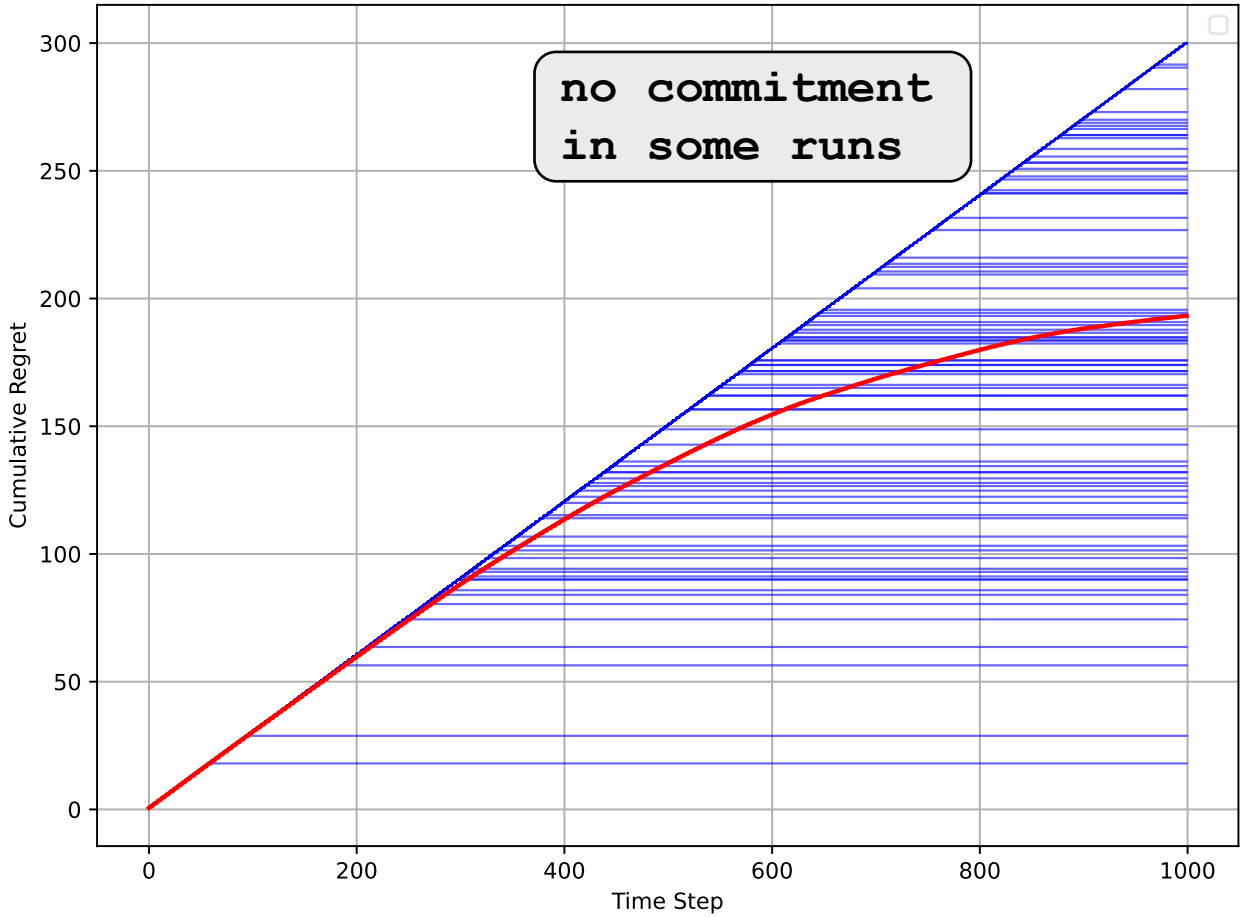
- Algorithm from View 1
  - All arms pulled equally often until commit time τ
  - Likely behavior: worse arms clearly separate out before competitive ones
  - Wasteful exploration on bad arms
  - Large regret

- Algorithm from View 2: Sequential Elimination
  - Initially, all arms are pulled in round robin
  - Algorithm proceeds in phases, eliminating one arm per phase
  - Length of each phase is random
  - Instantaneous regret in each phase reduces
  - Eventually left with only the best arm

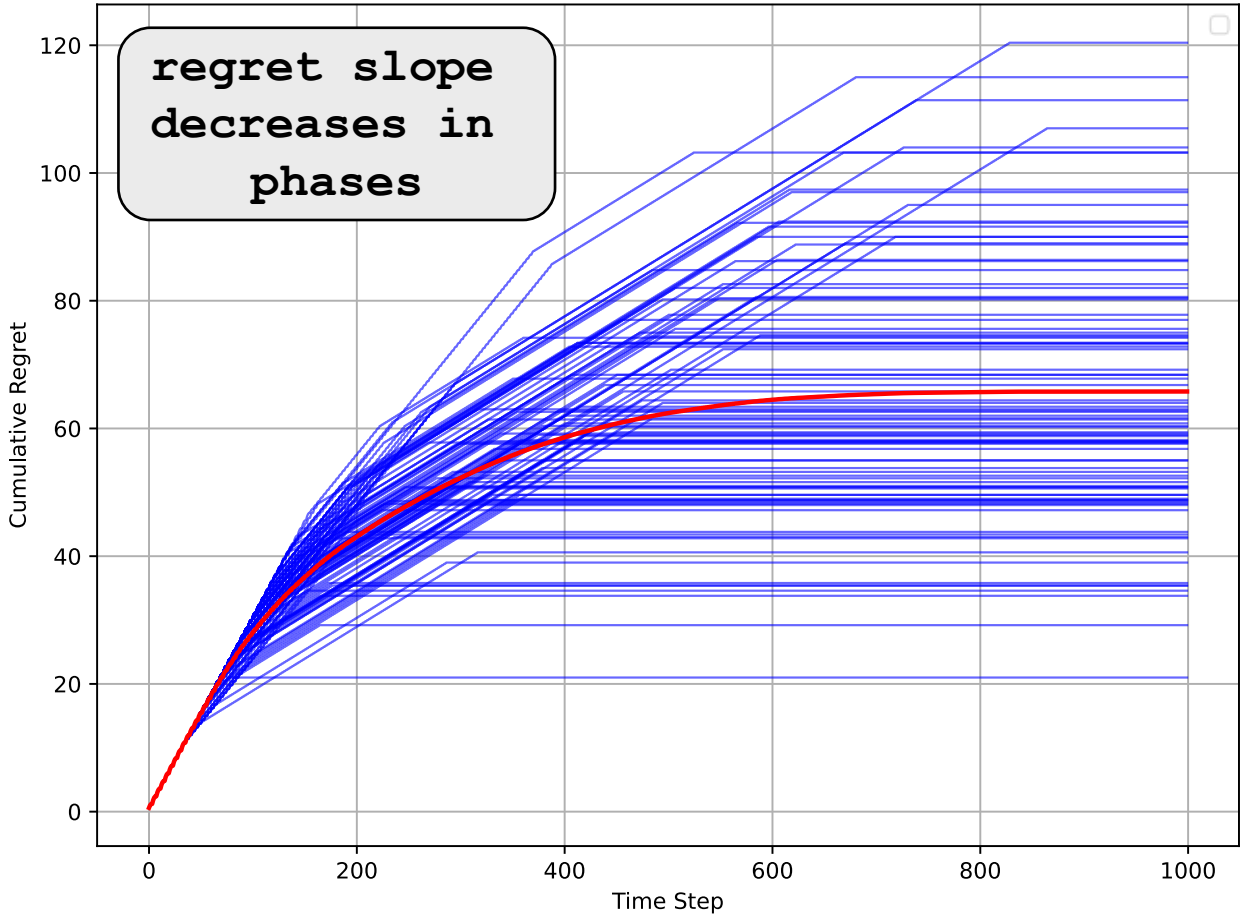
- Sequential Elimination is an extension of ETC to bandits with $k > 2$ arms

Adaptive ETC vs sequential elimination: regret curves

Adaptive Explore-Then-Commit



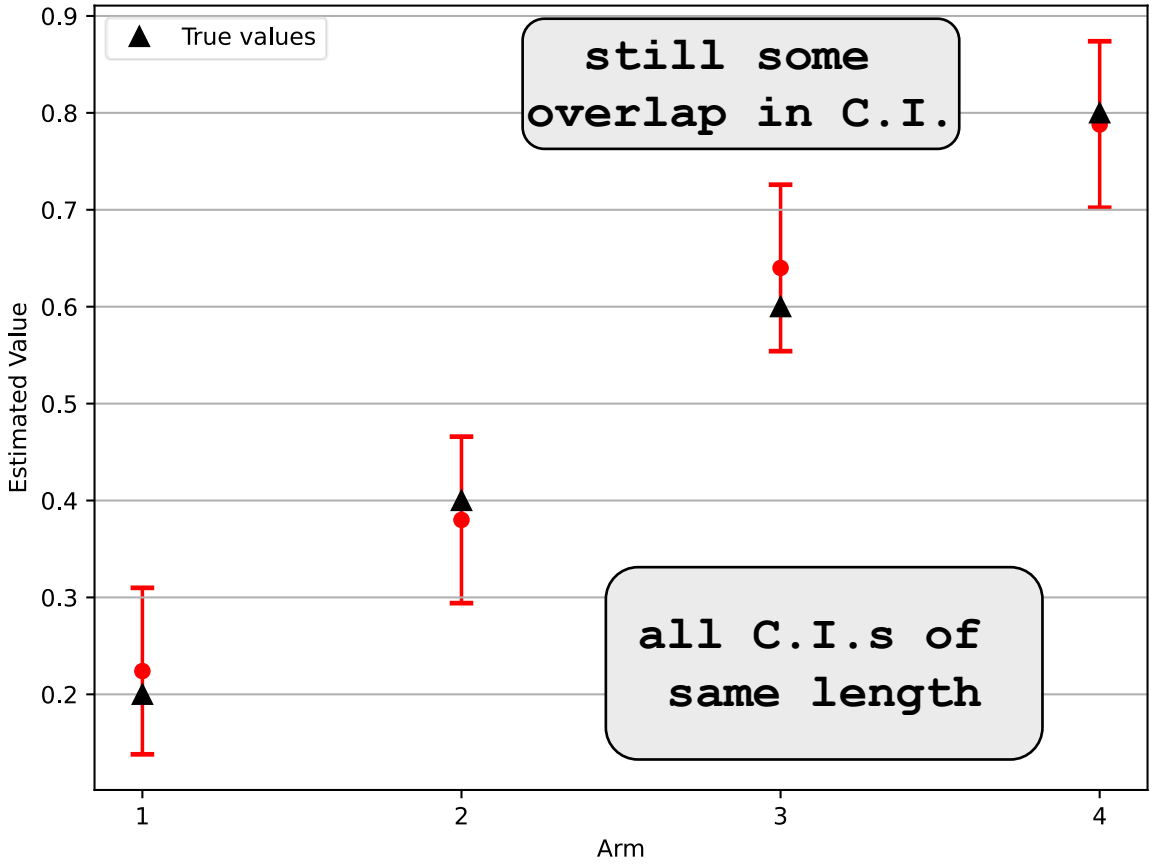
Sequential Elimination



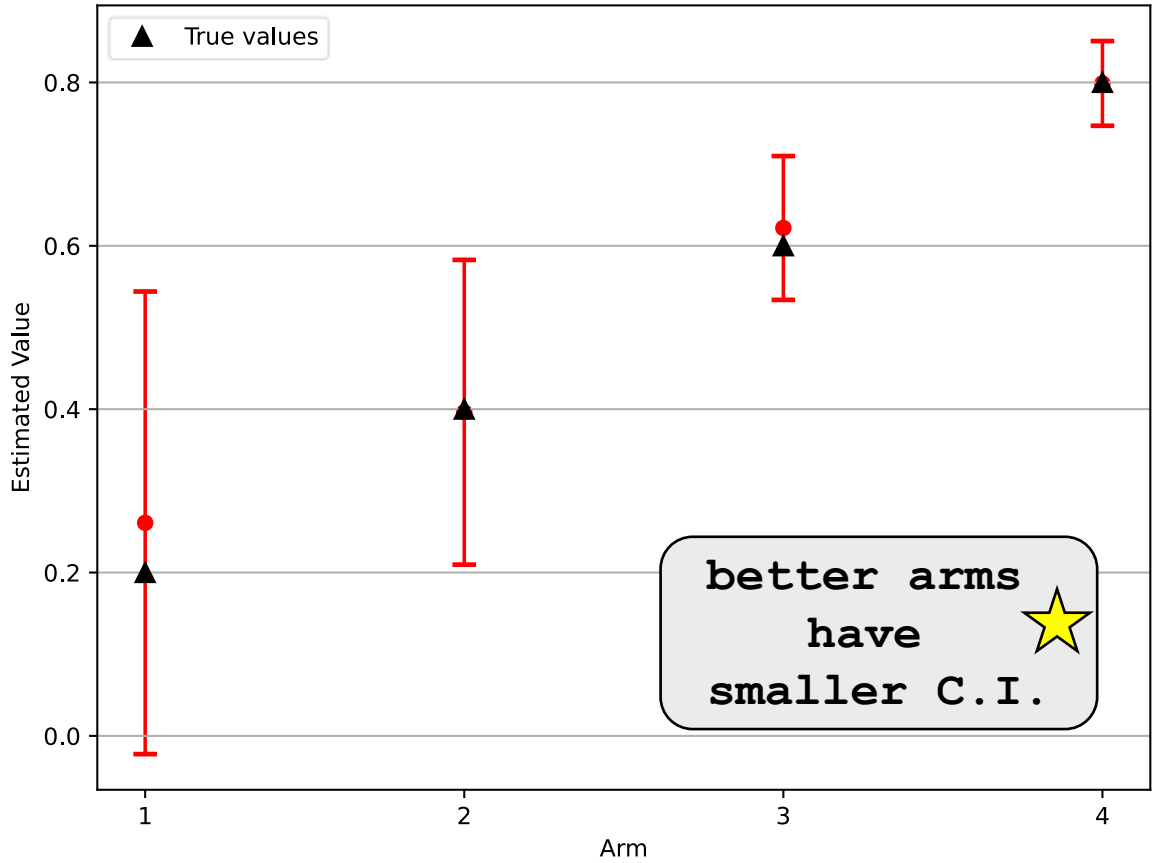
4-armed Bernoulli bandit with average reward [0.2, 0.4, 0.6, 0.8]

Adaptive ETC vs SE: confidence intervals

Adaptive Explore-Then-Commit



Sequential Elimination



4-armed Bernoulli bandit with average reward [0.2, 0.4, 0.6, 0.8]

Upper Confidence Bound (UCB) algorithm

- A popular algorithm for multi-arm bandits
- Uses confidence bounds, like flexible ETC and sequential elimination
- Leads to better regret bounds in practice

- Notation
  - estimate of mean reward of arm a at time t: $\hat{\mu}_a(t)$
  - number of pulls of arm a at time t: $T_a(t)$
  - upper confidence bound of arm a at time t: $UCB_a(t,\delta) = \hat{\mu}_a(t) + \sqrt{2\log(1/\delta)/T_a(t)}$

At time t refers to *after* t pulls of the arms

assume 1-subgaussian

initially ∞

UCB Algorithm

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

Choose $A_t = \arg\max_a UCB_a(t-1, \delta)$

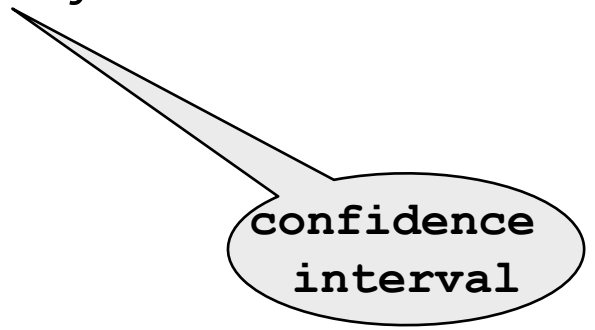
Observe X_t and update $UCB_a(t,\delta)$

Exploration-exploitation tradeoff in UCB

- Understanding $UCB_a(t, \delta)$
  - Empirical mean $\hat{\mu}_a(t)$ + Exploration bonus $\sqrt{\frac{2\log(1/\delta)}{T_a(t)}}$
- The algorithm pulls arms more often if they are
  - Promising
    - $\hat{\mu}_a(t)$ is large, or
  - Not well explored
    - $T_a(t)$ is small $\Rightarrow$ exploration bonus is large
- Why is this a reasonable strategy?
  - With high probability, $UCB_{a^*}(t, \delta) > \mu_{a^*}$ always
  - Arm $a \neq a^*$ is pulled $\Rightarrow UCB_a(t, \delta) > UCB_{a^*}(t, \delta)$
  - With enough pulls, $UCB_a(t, \delta) \approx \mu_a$ $(< \mu_{a^*} < UCB_{a^*}(t, \delta))$
  - Suboptimal arms pulled not too many times!

UCB versus ETC

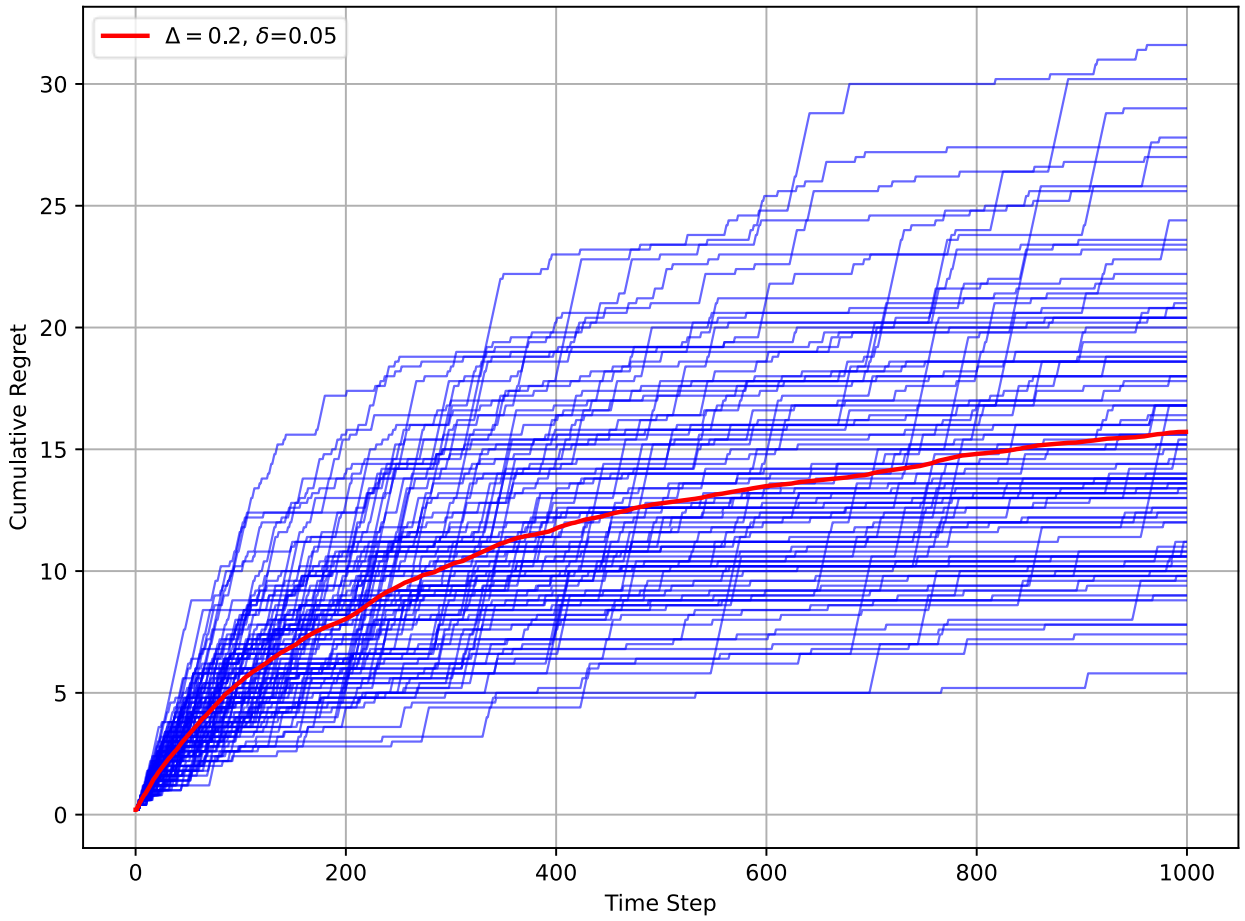
- UCB follows the Optimism in the Face of Uncertainty (OFU) principle
  - UCB plays assuming arm rewards are the highest they can reasonably be
  - Pretends $\mu_a(t) = UCB_a(t, \delta)$



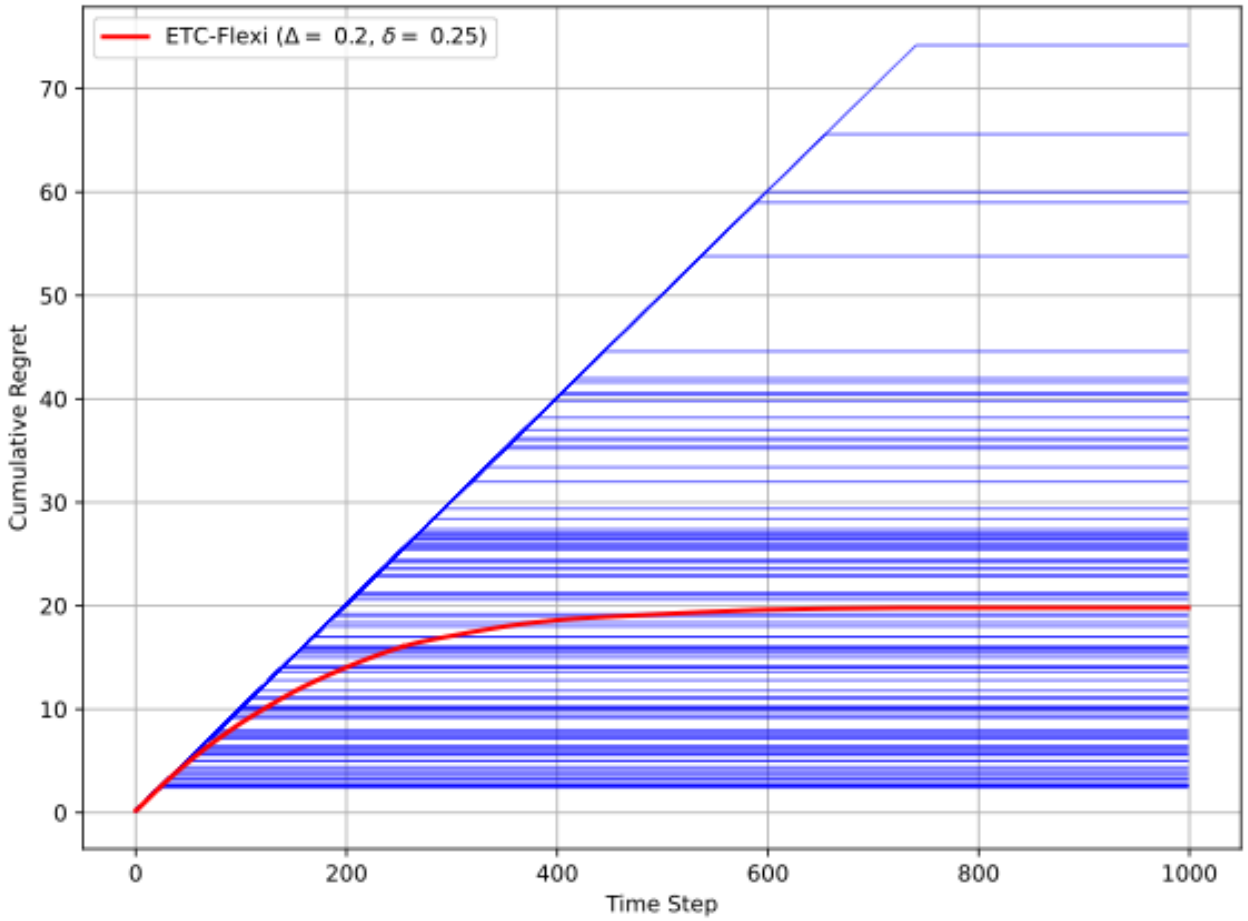
- ETC (and sequential elimination) is more conservative
  - Takes worst-case approach:
    - Arm rewards could be anything in the respective confidence intervals
  - Only commits best arm if 'sure' that it is best
    - Confidence intervals of best arm should dominate all others
  - Likewise, only eliminates bad arm if 'sure' it is not the best

- UCB has a soft transition from exploration to exploitation
- ETC has a hard transition: explore phase v/s commit phase
  - Difficult to correct if one makes a mistake

Regret performance of UCB

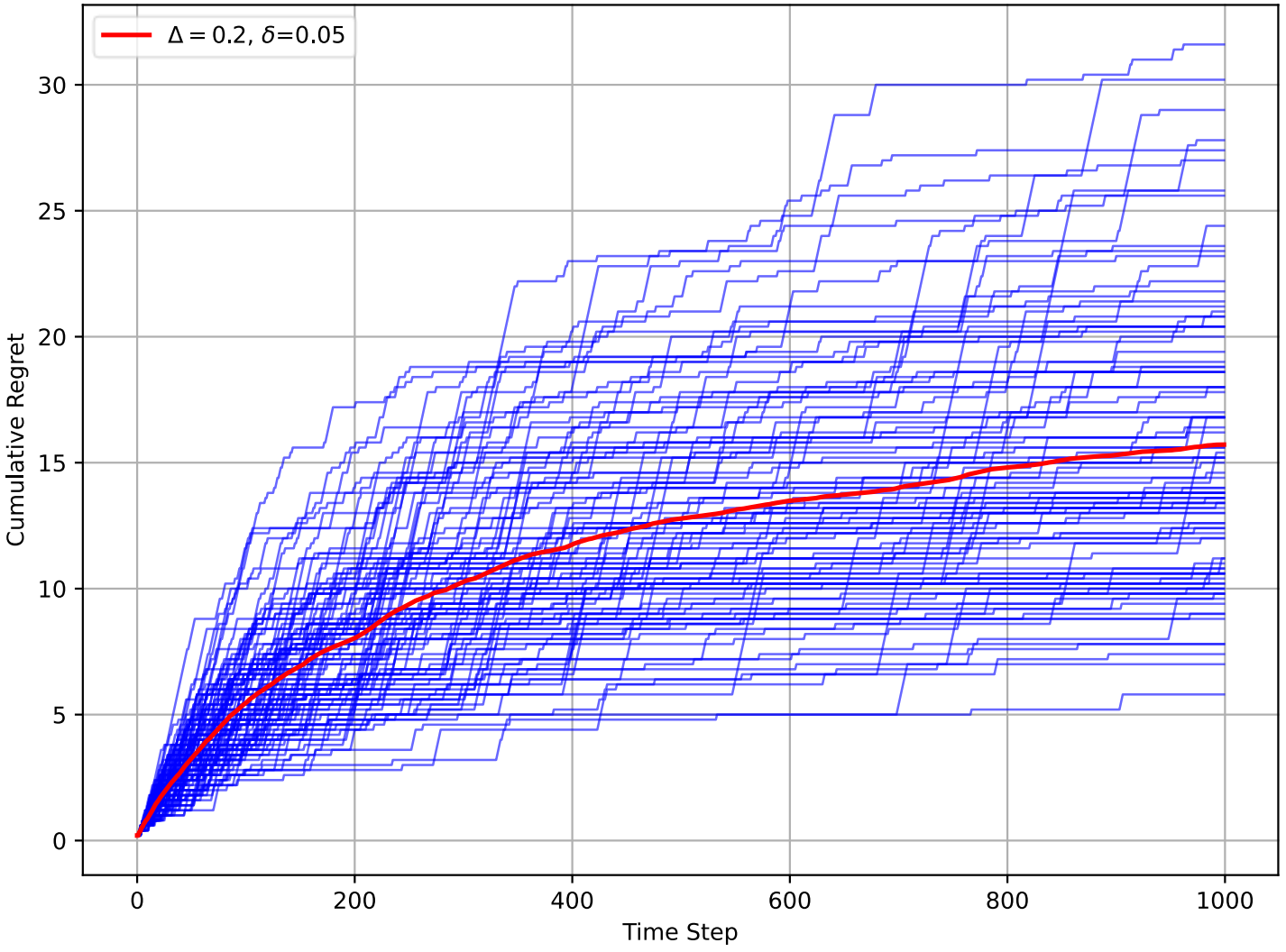


UCB Algorithm



Adaptive ETC

UCB vs ETC



Points to note:

- For each sample path, regret growth decays gently with time
- Contrast with ETC!
- Exploration never stops completely
- Sublinear mean regret
- Formula: $O(\sqrt{T})$? $O(\log T)$?

Regret bound for UCB

UCB Algorithm

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

Choose $A_t = \arg\max_a UCB_a(t-1, \delta)$

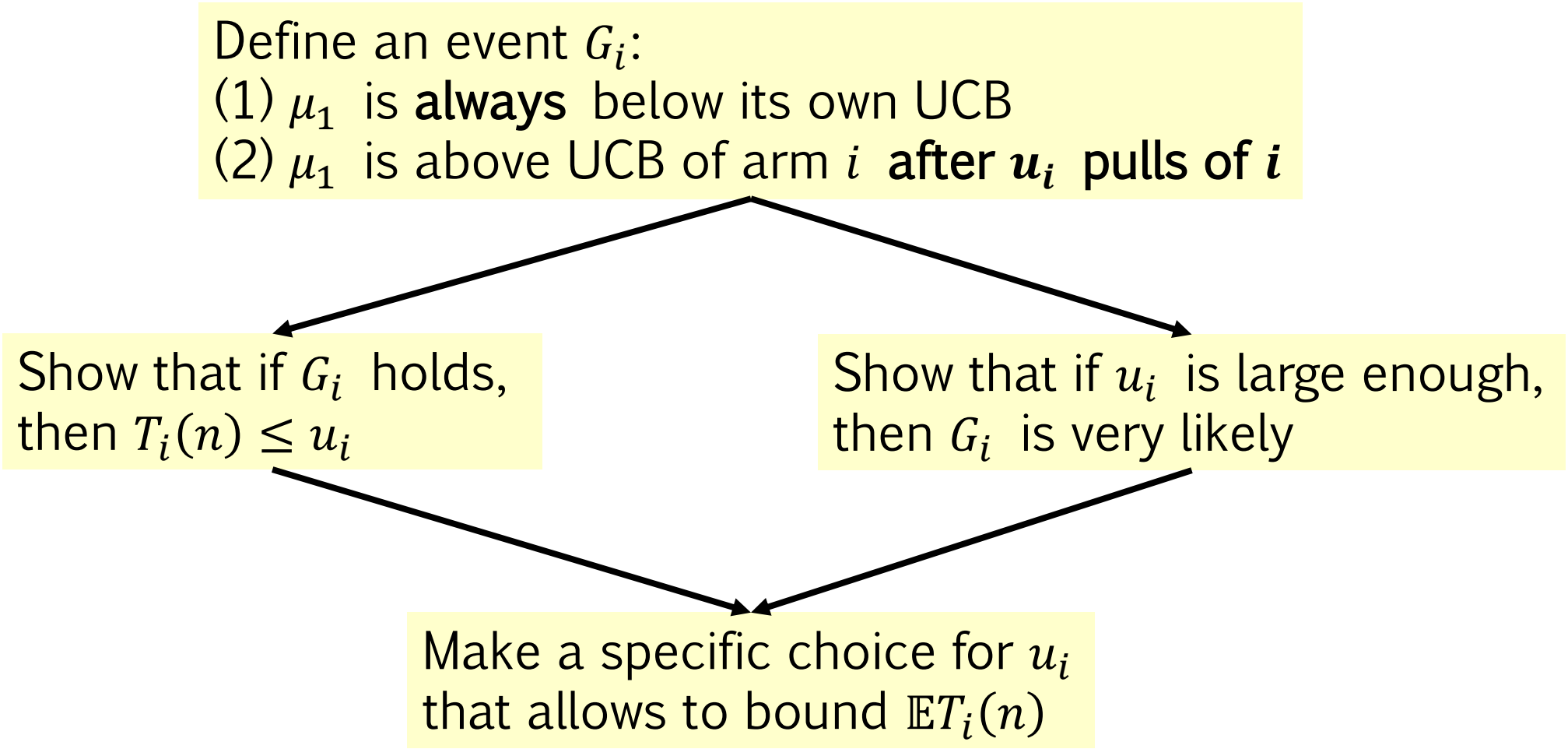
Observe X_t and update $UCB_a(t, \delta)$

- 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

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

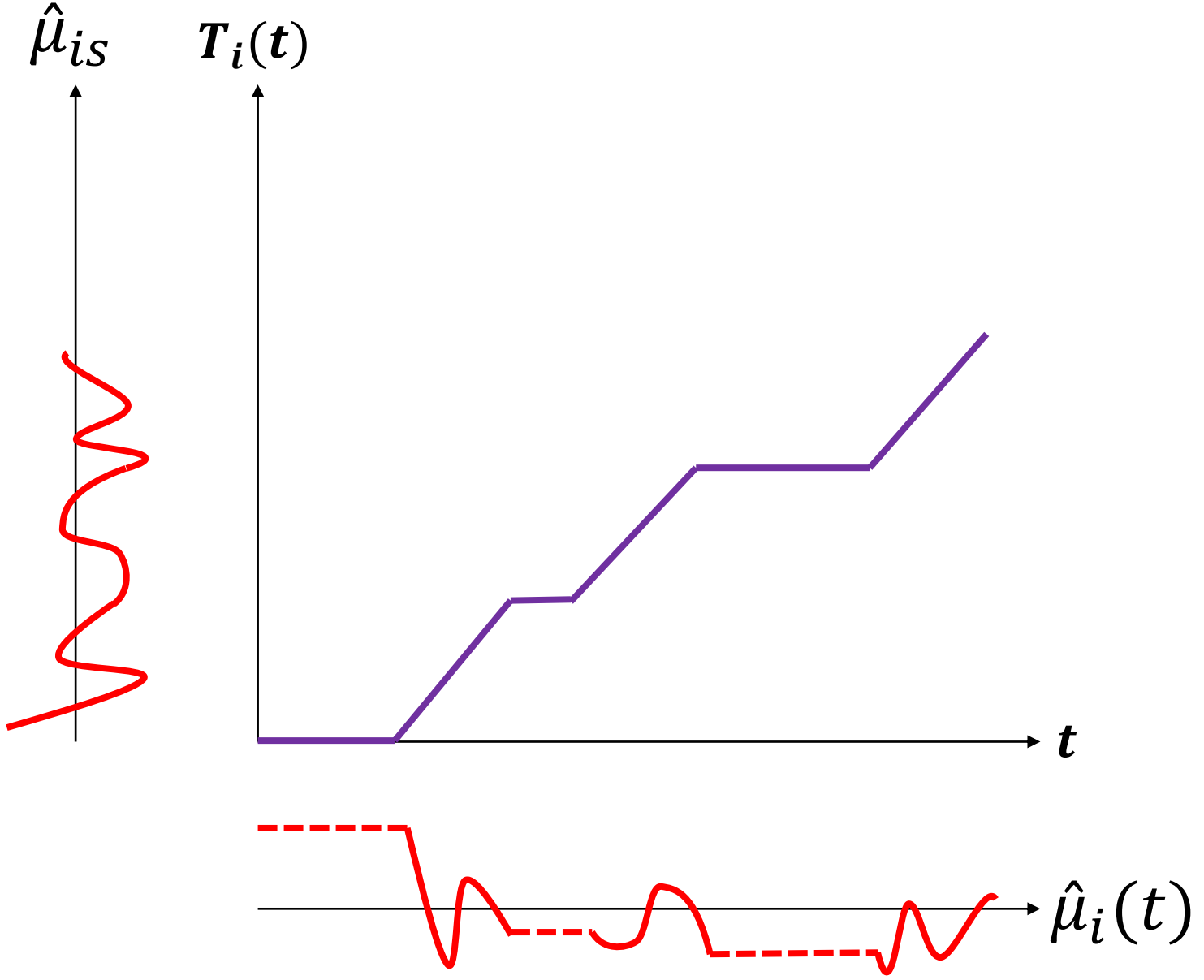
Proof of the UCB regret bound

- Outline:

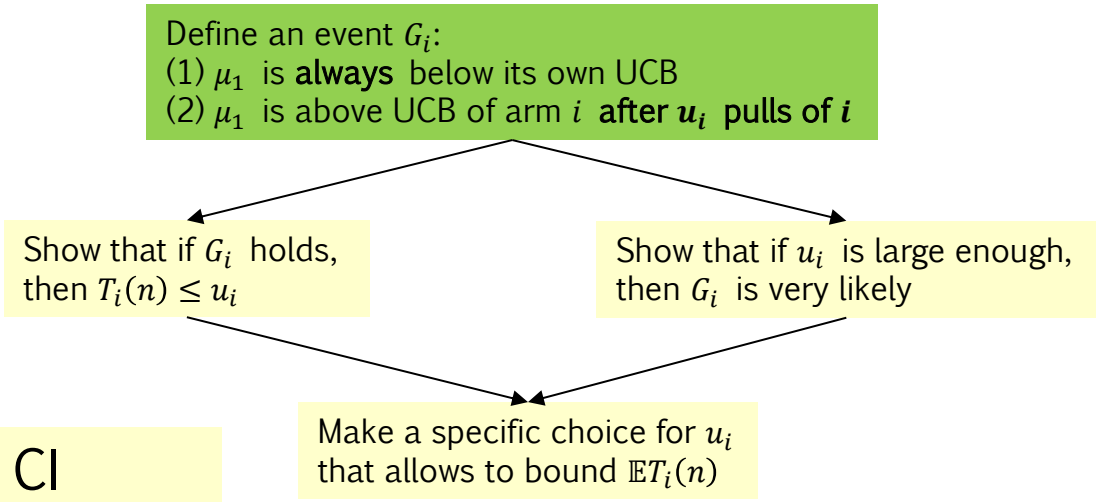


Notation: $\hat{\mu}_i(t)$ vs $\hat{\mu}_{is}$

- Estimator at time t: $\hat{\mu}_i(t)$
- Estimator after s pulls of arm i: $\hat{\mu}_{is}$
  - i.e., $\hat{\mu}_{is}$ is the average of s iid samples
- In other words: $\hat{\mu}_i(t) = \hat{\mu}_{iT_i(t)}$
- Remember: $T_i(t) \leq t$ for all t



Regret bound for UCB



- Good event G_i:
  - $\left\{\mu_1 < \min_{t\in[n]} UCB_1(t,\delta)\right\}$
  - $\cap$
  - $\left\{\hat{\mu}_{iu_i} + \sqrt{\frac{2}{u_i}\ln\left(\frac{1}{\delta}\right)} < \mu_1\right\}$

μ_1 is always inside the CI
(or more precisely, below UCB)

At some time t where $T_i(t) = u_i$,
arm i does not exceed μ_1

- Bound on $\mathbb{E}[T_i(n)]$:
  - $\mathbb{E}[T_i(n)] = \mathbb{E}[\mathbb{I}\{G_i\}T_i(n)] + \mathbb{E}[\mathbb{I}\{G_i^c\}T_i(n] \le u_i + \mathbb{P}(G_i^c)n$
  - We will need to choose u_i wisely, because there's a tradeoff:
    - u_i too large $\Rightarrow$ regret bound is too large
    - u_i too small $\Rightarrow$ $\mathbb{P}(G_i^c)$ is too large $\Rightarrow$ regret bound is too large

Regret bound for UCB

Define an event G_i:
(1) μ_1 is **always** below its own UCB
(2) μ_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 $\mathbb{E}T_i(n)$

- Goal: show that conditional on G_i, $T_i(n) \leq u_i$
- Proof: by contradiction: assume $T_i(n) > u_i$
  - Then arm i was pulled more than u_i times before n
  - So there must be a round t where arm i was pulled the $(u_i + 1)$th time
  - We use this to relate UCB_i to UCB_1:
    - $UCB_i(t-1,\delta) = \hat{\mu}_i(t-1) + \sqrt{\frac{2\ln\left(\frac{1}{\delta}\right)}{T_i(t-1)}}$

$$= \hat{\mu}_{iu_i} + \sqrt{\frac{2}{u_i}\ln\left(\frac{1}{\delta}\right)}$$

$$< \mu_1$$

$$< UCB_1(t-1,\delta)$$

$T_i(t-1) = u_i$

From definition of G_i

From definition of G_i

- This implies that we could not have pulled arm i at time t → contradiction!

Regret bound for UCB

- $G_i^c = \left\{\mu_1 \geq \min_{t\in[n]} UCB_1(t,\delta)\right\} \cup \left\{\hat{\mu}_{iu_i} + \sqrt{\frac{2}{u_i}\ln\left(\frac{1}{\delta}\right)} \geq \mu_1\right\}$
- Decompose the first event:
  - $\left\{\mu_1 \geq \min_{t\in[n]} UCB_1(t,\delta)\right\} \subset \left\{\mu_1 \geq \min_{s\in[n]} \hat{\mu}_{1s} + \sqrt{\frac{2}{s}\ln\left(\frac{1}{\delta}\right)}\right\}$

Note: $\min_{s\in[n]}$ is over n terms while $\min_{t\in[n]}$ is only over $T_i(n)$ terms – this can only lower the min

$$= \bigcup_{s\in[n]} \left\{\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2}{s}\ln\left(\frac{1}{\delta}\right)}\right\}$$

- Now we bound using subgaussian assumption and concentration results:
  - $\mathbb{P}\left[\mu_1 \geq \min_{t\in[n]} UCB_1(t,\delta)\right] \leq \mathbb{P}\left[\cup_{s\in[n]} \left\{\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2}{s}\ln\left(\frac{1}{\delta}\right)}\right\}\right]$

$$\leq \sum_{s=1}^{n} \mathbb{P}\left[\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2}{s}\ln\left(\frac{1}{\delta}\right)}\right] \leq n\delta$$

Regret bound for UCB

Define an event G_i:
(1) μ_1 is **always** below its own UCB
(2) μ_1 is above UCB of arm i **after u_i pulls of i**

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

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 $\mathbb{E}T_i(n)$

- To bound the second event, we need to make an assumption on u_i:
  - Assume u_i is large enough so that $\Delta_i - \sqrt{\frac{2\ln\left(\frac{1}{\delta}\right)}{u_i}} \ge c\Delta_i$ for some constant $0 < c < 1$
  - $$\mathbb{P}\left[\hat{\mu}_{iu_i} + \sqrt{\frac{2}{u_i}\ln\left(\frac{1}{\delta}\right)} \ge \mu_1\right] = \mathbb{P}\left[\hat{\mu}_{iu_i} - \mu_i \ge \Delta_i - \sqrt{\frac{2}{u_i}\ln\left(\frac{1}{\delta}\right)}\right]$$
    $$\le \mathbb{P}\left[\hat{\mu}_{iu_i} - \mu_i \ge c\Delta_i\right]$$
    $$\le \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)$$

$\mu_1 = \mu_i + \Delta_i$

Chernoff bound
for subgaussian sums:
u_i is the number of terms in $\hat{\mu}_{iu_i}$

- Combining the two bounds:
  - $\mathbb{P}[G_i^c] \le n\delta + \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)$

Regret bound for UCB

Define an event G_i:
(1) μ_1 is **always** below its own UCB
(2) μ_1 is above UCB of arm i **after u_i pulls of i**

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

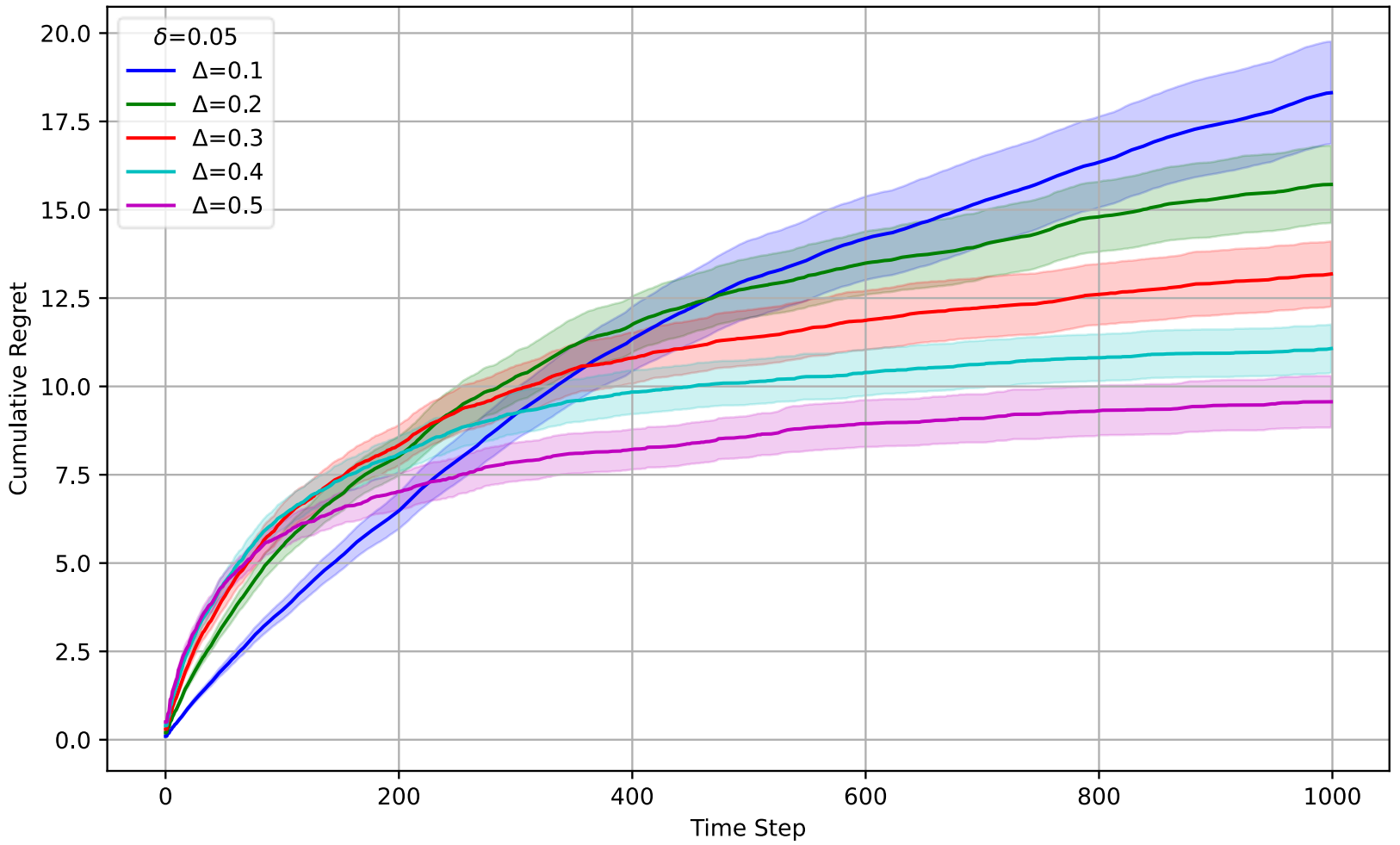
Show that if u_i is large enough, then G_i is very likely

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

- Back to $\mathbb{E}[T_i(n)] \le u_i + \mathbb{P}(G_i^c)n$
$$\le u_i + n\left(n\delta + \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)\right)$$
- Set u_i as small as possible given $\Delta_i - \sqrt{\frac{2\ln\left(\frac{1}{\delta}\right)}{u_i}} \ge c\Delta_i$
  - $u_i = \left\lceil \frac{2\ln(1/\delta)}{(1-c)^2\Delta_i^2} \right\rceil$
  - Note: no guarantee that this is $\le n$, but if it is not, bound on $\mathbb{E}[T_i(n)]$ holds trivially
- Recall: $\delta = 1/n^2$ → $\exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right) = \exp\left(-\frac{2\ln(n)c^2}{(1-c)^2}\right) = n^{-\frac{2c^2}{(1-c)^2}}$
- $\mathbb{E}[T_i(n)] \le \left\lceil \frac{2\ln(1/\delta)}{(1-c)^2\Delta_i^2} \right\rceil + 1 + n^{1-2c^2/(1-c)^2}$
  → choose c such that last term does not blow up, e.g., $c = 0.5$
- $\mathbb{E}[T_i(n)] \le 3 + \frac{16\ln(n)}{\Delta_i^2}$

UCB regret plots: variation with Δ

Bernoulli rewards with means $[0.5 - \frac{\Delta}{2}, 0.5 + \frac{\Delta}{2}]$



- Sublinear regret for all values of Δ, given fixed hyperparameter δ
- Cumulative regret increases inversely with Δ
- Slope gradually reduces with time, signifying shift from exploration to exploitation
- Initial regret slope higher for large Δ, due to expensive exploration

UCB versus ETC

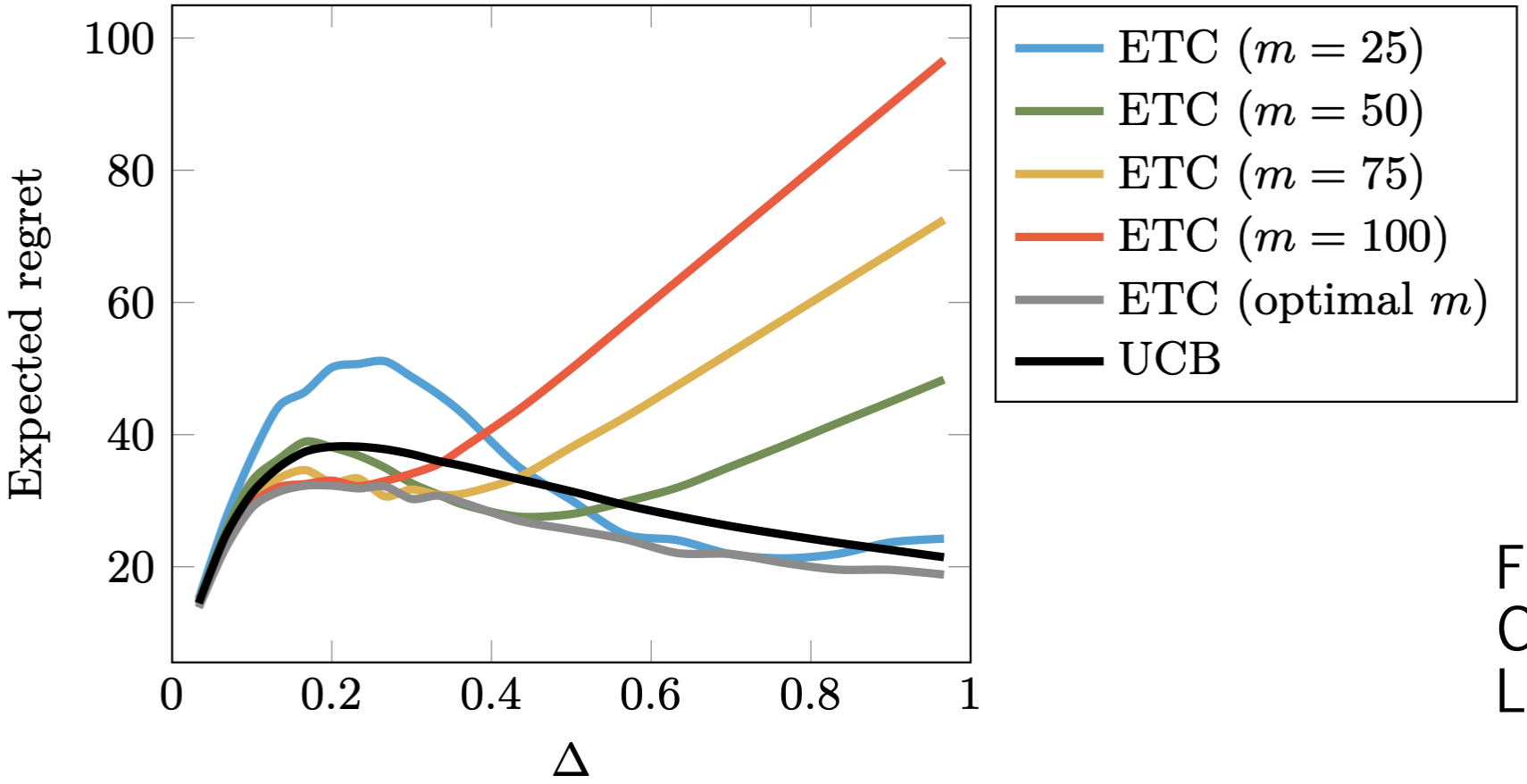


Figure from Chapter 7, L&S

UCB performs similar to ETC, without knowing optimal m (suboptimality gap)

Summary

- First, we saw
  - Explore-Then-Commit Algorithm with flexible stopping time
  - Chooses exploration phase adaptively, based on data
  - Gives sublinear regret
  - Extends to sequential elimination for multiple arms

- Main takeaway: UCB Algorithm
  - Quantifies uncertainty with confidence intervals
  - Acts optimistically in the face of uncertainty
  - Smooth transition from exploration to exploitation
  - Derived sublinear regret bounds
  - Reading L&S: Chapter 7 (especially Theorem 7.1)

- Coming Up Next Week:
  - Alternate proof of regret bound (Theorem 7.2)
  - Other bandit algorithms (Thompson Sampling (Chapter 36))
  - Other bandit formulations (Best-Arm Identification (Chapter 33))