Concentration of Measure and Explore-then-Commit Algorithm

Principles of Online Decision-Making (CS-303)

Prof. Matthias Grossglauser

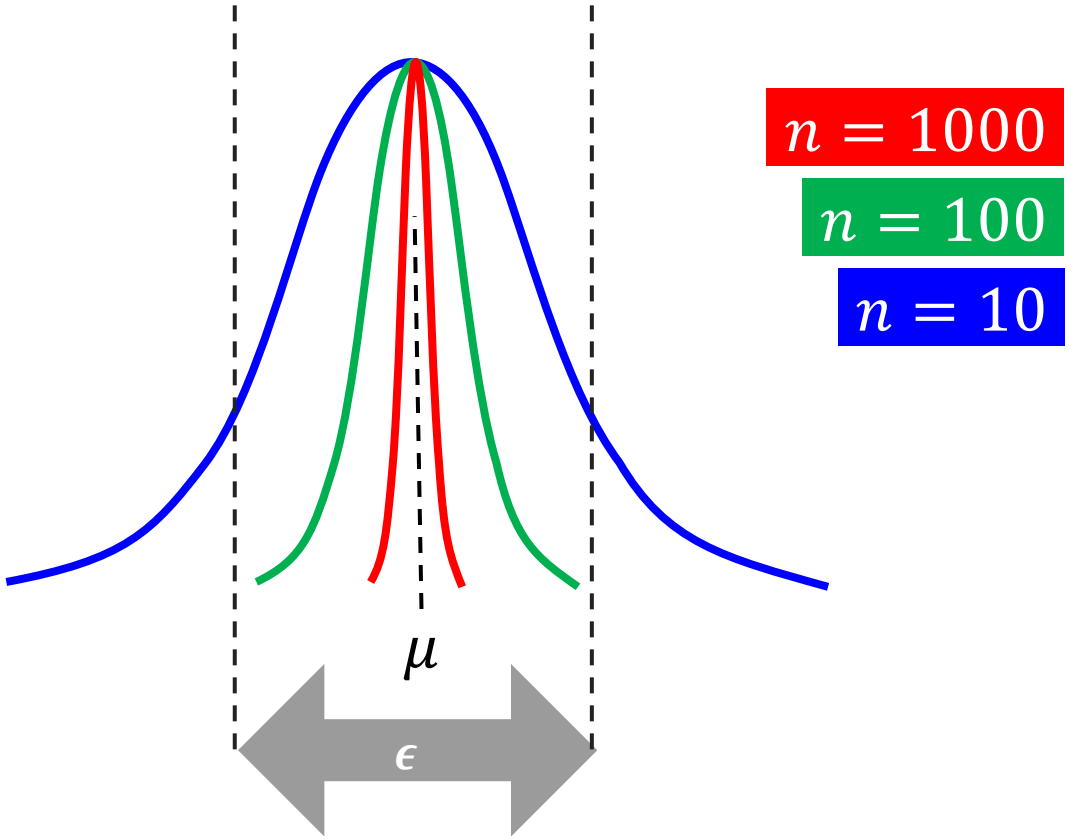
Information and Network Dynamics (INDY) lab
School of Computer and Communication Sciences (I&C)
EPFL

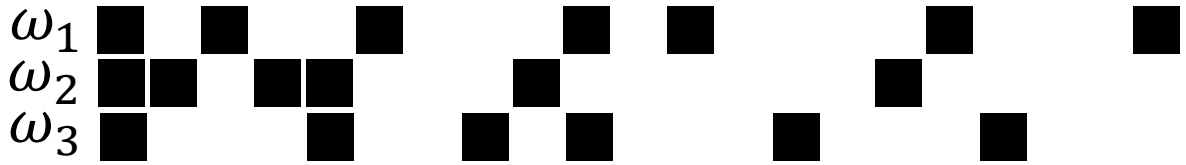
Sums of i.i.d. random variables

- To analyze bandit algorithms, we have to deal with sums of random regrets
- $X_1, X_2, X_3, \ldots, X_n$: i.i.d. whose mean and variance exist
  - $\mu = \mathbb{E}X$
  - $\sigma^2 = \mathbb{V}X = \mathbb{E}[(X - \mathbb{E}X)^2]$
- Estimator for μ:
  - $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \stackrel{\text{def}}{=} \frac{S_n}{n}$
  - $\mathbb{E}[\hat{\mu}] = \mu$: unbiased estimator
  - $\mathbb{V}[\hat{\mu}] = \mathbb{E}[(\hat{\mu} - \mu)^2] = \frac{\sigma^2}{n}$: estimator variance decreases

Large n limit: law of large numbers

- Law of large numbers
  - Weak law
    - $\lim_{n\to\infty} \mathbb{P}[|S_n - n\mu| \geq n\epsilon] \to 0$
  - Strong law
    - $\mathbb{P}\left[\lim_{n\to\infty} \frac{S_n}{n} = \mu\right] = 1$
- Side note:
  - Weak law: convergence in probability
  - Strong law: almost sure convergence
  - Example of a sequence (not iid) that converges to 0 in probability but not almost surely:
    - $Y_n = \begin{cases} 1 & \text{randomly with prob. } 1/n \\ 0 & \text{otherwise} \end{cases}$



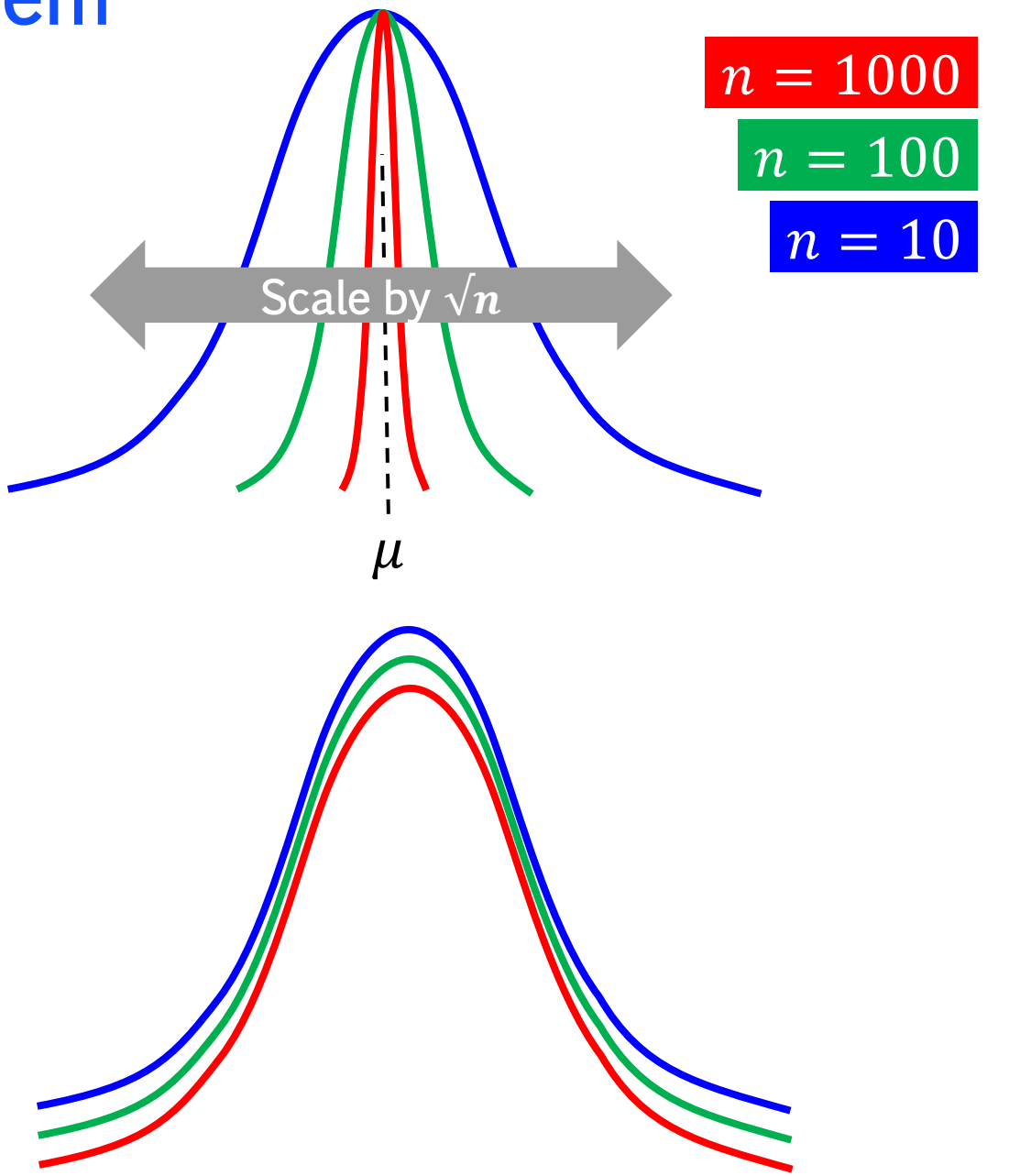


Large n limit: central limit theorem

- Sum: $S_n = \sum_{i=1}^{n} X_i$
- Rescale $\hat{\mu} - \mu$ by the inverse of its standard deviation $(1/\sqrt{n})$:

  $$\mathbb{P}\left[S_n - n\mu \geq \sqrt{n}\epsilon\right] \to \mathbb{P}\left[\mathcal{N}\left(0, \sigma^2\right) \geq \epsilon\right]$$

- This is the primary reason why the Gaussian law is so central
  - Plus stability under addition



Markov and Chebyshev inequalities

- Markov: $\mathbb{P}(|X| \geq \epsilon) \leq \frac{\mathbb{E}|X|}{\epsilon}$
  - Set $Y = |X|$, then $\mathbb{E}Y = \int_{y=0}^{\infty} yf(y)dy \geq \int_{y=\epsilon}^{\infty} \epsilon f(y)dy = \epsilon\mathbb{P}[Y \geq \epsilon]$
- Chebyshev: $\mathbb{P}(|X - \mathbb{E}X| \geq \epsilon) \leq \frac{\mathbb{V}X}{\epsilon^2}$
  - Obtained by applying Markov inequality to $(X - \mathbb{E}X)^2$
- Applied to the empirical mean:
  - $\mathbb{P}(|\hat{\mu} - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$
  - Pro: nice bound, only need mean and variance
  - Con: the bound is loose: decreases as $1/n$
    - We need something much stronger for the analysis of bandit algorithms

Central Limit Theorem (CLT)

- $S_n = \sum_{i=1}^{n}(X_i - \mu)$
- CLT: If mean and variance exist, then $\frac{S_n}{\sqrt{n\sigma^2}} \to \mathcal{N}(0,1)$
- Let $Z = \mathcal{N}(0,1)$, and let us bound $\mathbb{P}(Z \geq z) = \int_z^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$
  - This CDF has no closed-form solution → bound
  - $\int_z^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \leq \frac{1}{z\sqrt{2\pi}} \int_z^\infty x \exp\left(-\frac{x^2}{2}\right) dx = \sqrt{\frac{1}{2\pi z^2}} \exp\left(-\frac{z^2}{2}\right)$
  - $\mathbb{P}(\hat{\mu} - \mu \geq \epsilon) = \mathbb{P}\left[\frac{S_n}{\sqrt{\sigma^2 n}} \geq \epsilon\sqrt{\frac{n}{\sigma^2}}\right] \to \mathbb{P}\left[Z \geq \epsilon\sqrt{\frac{n}{\sigma^2}}\right] \leq \sqrt{\frac{\sigma^2}{2\pi n\epsilon^2}} \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)$
- Pro: in general, much tighter than Chebyshev (exponential in n vs $1/n$)
- Con: this is an asymptotic result, we cannot bound the probability for a finite n
  - We will need to pay for this limitation by making additional assumptions

Subgaussian random variable

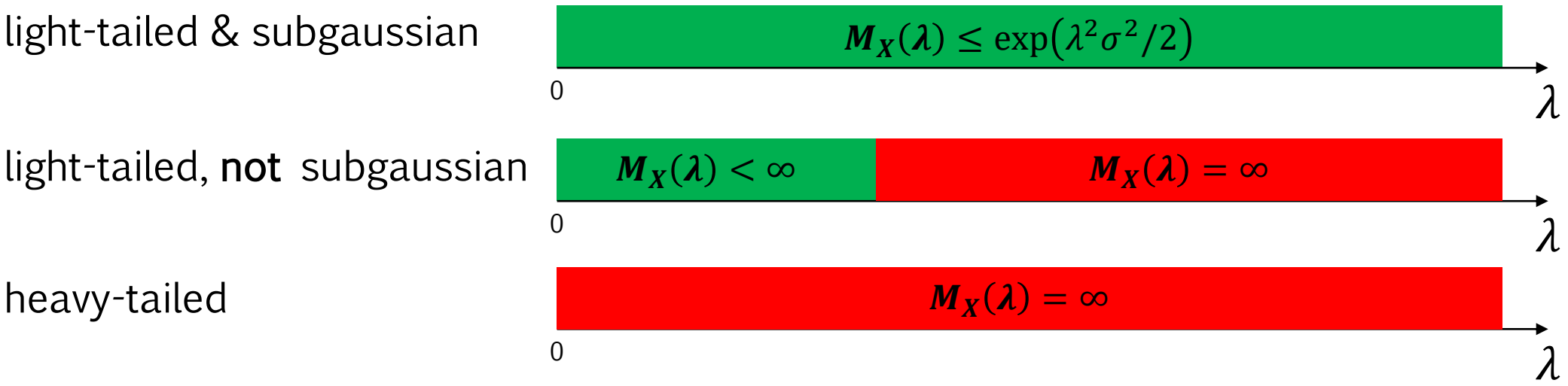
- Def: a random variable X is σ-subgaussian if for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\lambda X}\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$$

  - Note: If X is σ-subgaussian, then it is also σ'-subgaussian for any $\sigma' > \sigma$
- Def: a random variable X is σ-subgaussian if for all $\lambda \in \mathbb{R}$,

$$\psi_X(\lambda) = \log M_X(\lambda) \leq \frac{1}{2}\lambda^2\sigma^2$$

  - $M_X(\lambda)$: moment-generating function (MGF)



Properties of subgaussian random variables

- Assume X is σ-subgaussian
  - $\mathbb{E}X = 0$ and $\mathbb{V}X \leq \sigma^2$
    - Note: we sometimes abuse notation and refer to subgaussian RVs that are not zero mean → subtract mean
  - αX is $|\alpha|\sigma$-subgaussian
- Assume X_1 is σ_1-subgaussian and X_2 is σ_2-subgaussian, $X_1 \perp X_2$
  - $X_1 + X_2$ is $\sqrt{\sigma_1^2 + \sigma_2^2}$-subgaussian

Tail of a subgaussian random variable

- Thm: For X σ-subgaussian,

$$\mathbb{P}[X \geq \epsilon] \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$$

- Proof:
  - For some $\lambda > 0$
  - $$\mathbb{P}[X \geq \epsilon] = \mathbb{P}[\exp(\lambda X) \geq \exp(\lambda\epsilon)]$$
    $$\leq \mathbb{E}[\exp(\lambda X)]\exp(-\lambda\epsilon)$$ Markov's inequality
    $$\leq \exp\left(\frac{\lambda^2\sigma^2}{2} - \lambda\epsilon\right)$$ By definition of subgaussianity
  - Set $\lambda = \epsilon/\sigma^2$
- Corresponding inequality for left tail → combined: $\mathbb{P}[|X| \geq \epsilon] \leq 2\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$

Confidence intervals

- Inverting this relationship → confidence intervals for X
  - Set some small probability (error tolerance) δ
  - Two-sided: by union bound $\mathbb{P}[A \cup B] \leq \mathbb{P}[A] + \mathbb{P}[B]$ → $\mathbb{P}[|X| \geq \epsilon] \leq 2\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$
  - Then $\mathbb{P}\left[X \geq \sqrt{2\sigma^2 \ln(1/\delta)}\right] \leq \delta$, or $\mathbb{P}\left[|X| \geq \sqrt{2\sigma^2 \ln(2/\delta)}\right] \leq \delta$
- So with overwhelming probability at least $1-\delta$, X concentrates in this confidence interval:

$$\left(-\sqrt{2\sigma^2 \ln\left(\frac{2}{\delta}\right)}, +\sqrt{2\sigma^2 \ln\left(\frac{2}{\delta}\right)}\right)$$

- We will apply such confidence intervals to characterize the estimators of per-arm reward in different bandit algorithms

Tail of a subgaussian random variable

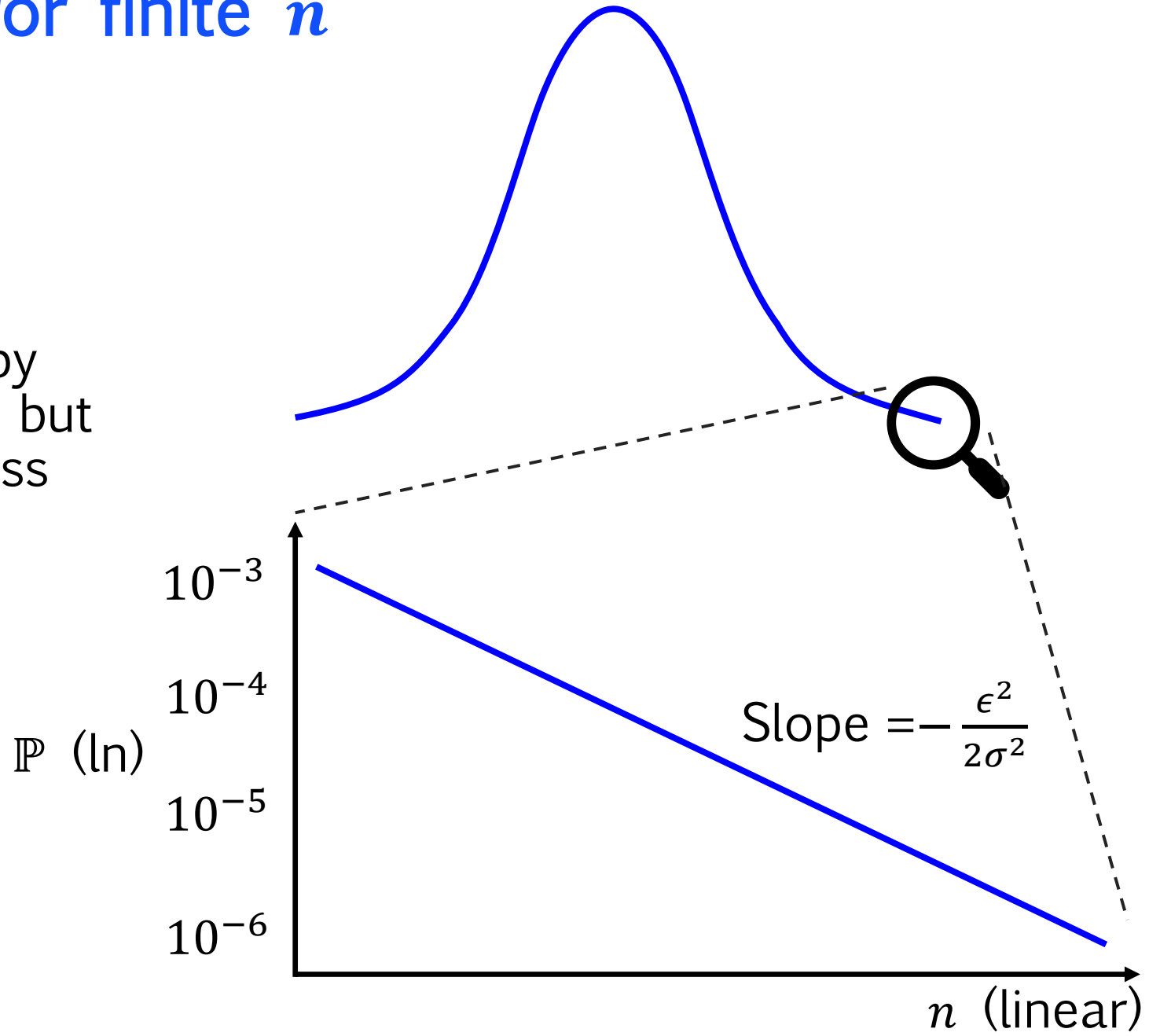
- Applied to sums of independent RVs:
  - Thm: assume $X_i - \mu$ are independent σ-subgaussian RVs. For any $\epsilon > 0$

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

  - Proof:
    - $\hat{\mu} - \mu = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)$ is $\sigma/\sqrt{n}$-subgaussian
    - Apply preceding theorem
- Confidence intervals:
  - $\mathbb{P}\left[\mu \geq \hat{\mu} + \sqrt{\frac{2\sigma^2 \ln(1/\delta)}{n}}\right] \leq \delta$ and $\mathbb{P}\left[\mu \leq \hat{\mu} - \sqrt{\frac{2\sigma^2 \ln(1/\delta)}{n}}\right] \leq \delta$
  - Two-sided: $\mathbb{P}\left[\mu \in \left(\hat{\mu} - \sqrt{\frac{2\sigma^2 \ln(2/\delta)}{n}}, \hat{\mu} + \sqrt{\frac{2\sigma^2 \ln(2/\delta)}{n}}\right)\right] \geq 1 - \delta$

Large deviation bound for finite n

- $\mathbb{P}[\hat{\mu} \geq \mu + \epsilon] \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)$
- $\mathbb{P}[S_n - n\mu \geq n\epsilon] \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)$
- Similar bound as suggested by CLT (without being a bound), but only for a (much smaller) class of variables



Examples of subgaussian random variables

- Gaussian is subgaussian:
  - $M_X(\lambda) = \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$
  - $X \sim \mathcal{N}(0, \sigma^2)$ → σ-subgaussian
- Bounded is subgaussian
  - $X \in [a, b]$ (a.s.) → $(b-a)/2$-subgaussian
  - $|X| \leq a$ → a-subgaussian
- Exponential is **not** subgaussian (tail is not light enough!)
  - $X \sim \mathrm{Exp}(1/\mu)$
  - $M_X(\lambda) = \int_0^\infty \mu^{-1} e^{-\frac{x}{\mu}} e^{\lambda x} dx = \begin{cases} (1-\lambda\mu)^{-1} & \lambda < 1/\mu \\ \infty & \text{otherwise} \end{cases}$

The Explore-then-Commit (ETC) algorithm

- Recall secretary problem: pure exploration for fixed # of steps, then pure exploitation
- Can try a similar policy for bandits: invest some predetermined number of samples to estimate mean reward per arm, then play best arm until n
- Definitions:
  - Estimated reward for arm i: $\hat{\mu}_i(t) = \frac{1}{T_i(t)} \sum_{s=1}^{t} \mathbb{I}\{A_s = i\} X_s$
  - Number of pulls of arm i: $T_i(t) = \sum_{s=1}^{t} \mathbb{I}\{A_s = i\}$
- Only parameter to set: number of exploration steps
  - m per arm, mk overall
- The algorithm is generic, but the analysis will need to assume subgaussian rewards for all arms

The Explore-then-Commit (ETC) algorithm

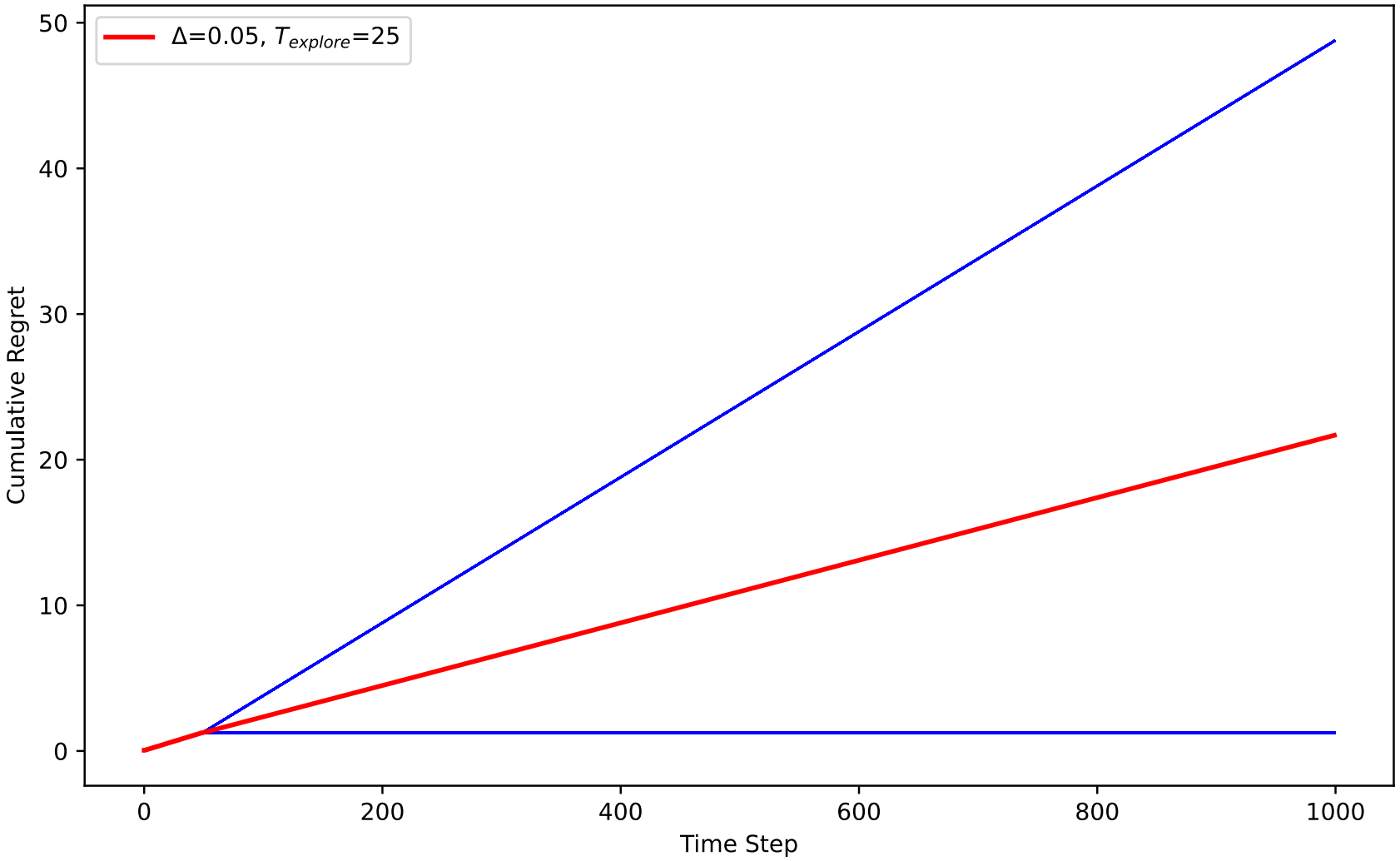
Explore-then-commit (m):

for t = 1…n:

$$A_t = \begin{cases} (t \bmod k) + 1 & t \le mk \\ \arg\max \hat{\mu}_i(mk) & t > mk \end{cases}$$

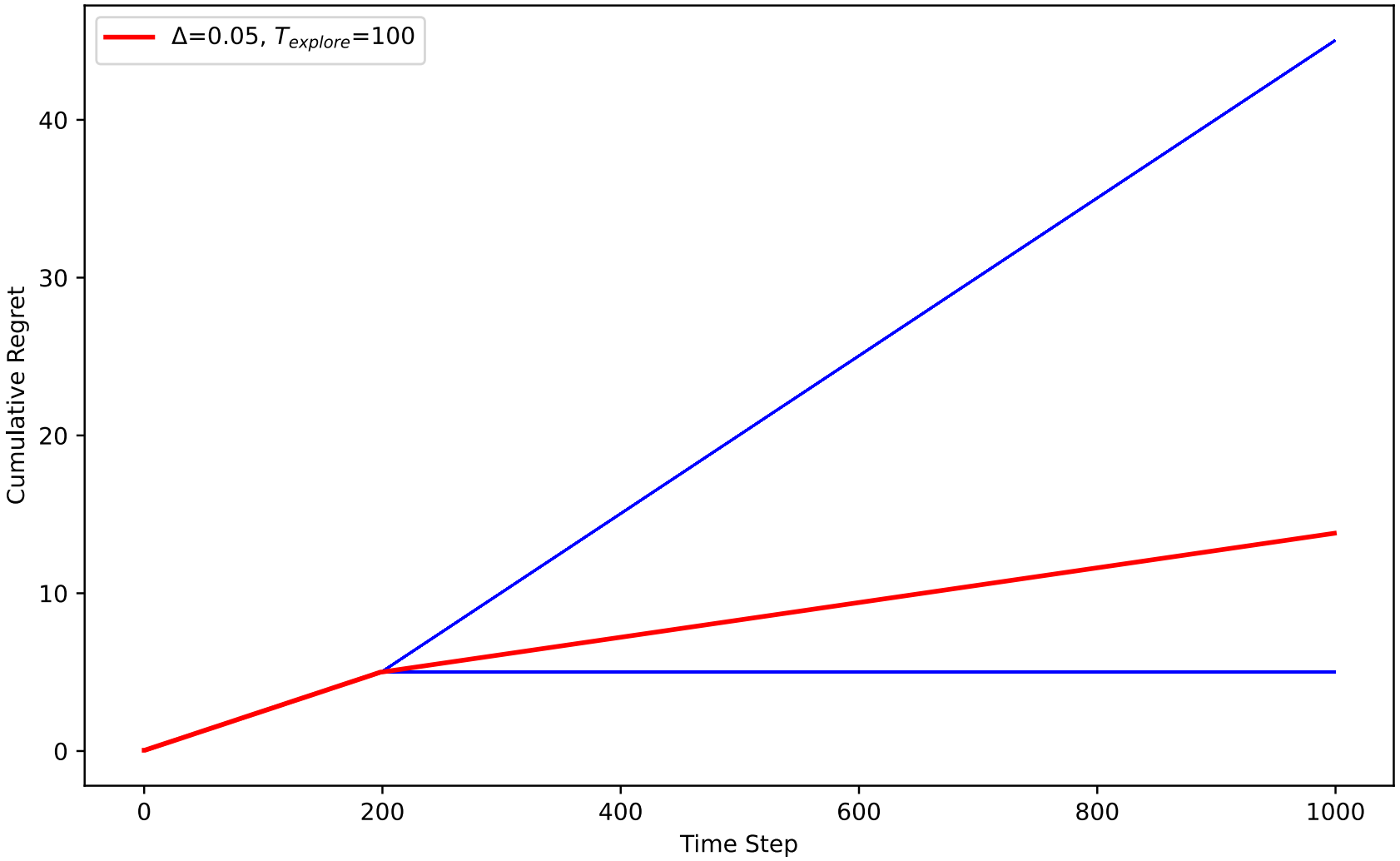
- m: # of explore-pulls per arm
- km: total duration of explore phase
- Note: order of pulls in explore phase does not matter, it just needs to be m per arm
  - Could do all arm 1 back-to-back, then arm 2, etc.
- Let's look at some simulations first:
  - Bernoulli bandit with $k = 2$

Small $\Delta = 0.05$, small $m = 25$

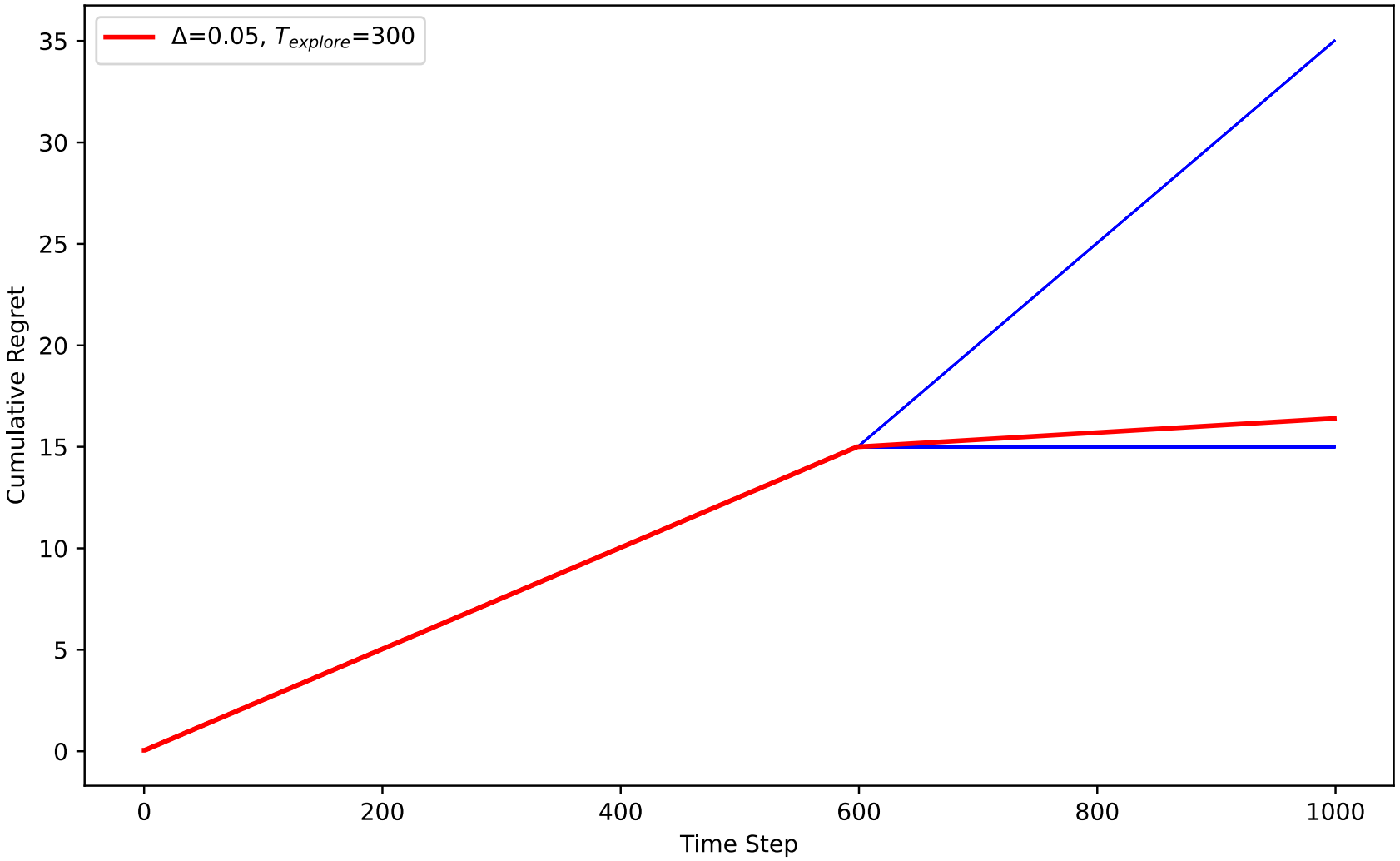


- Blue:
  $$\sum_{a \in \mathcal{A}} \Delta_a T_a(t)$$
  = trajectory-dependent reward
- Red:
  $$R_n = \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E} T_a(t)$$
  = expected reward

Small $\Delta = 0.05$, medium $m = 100$



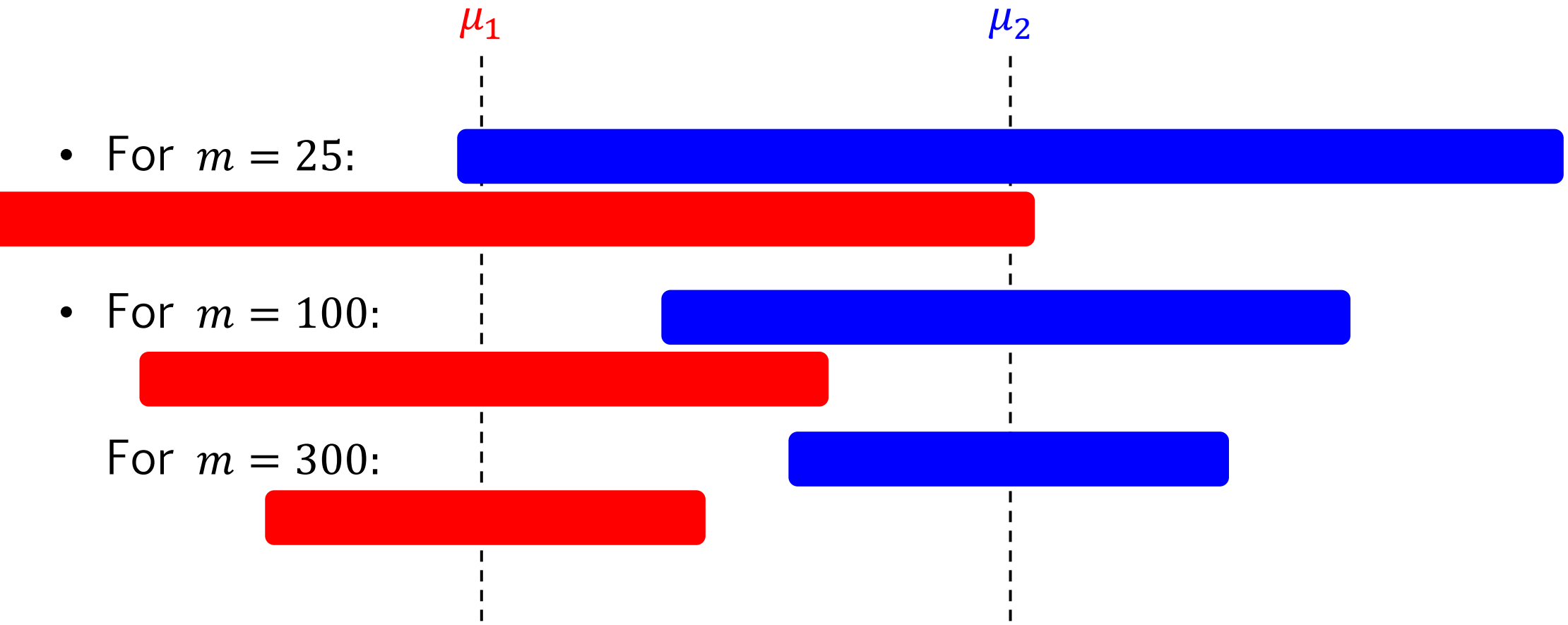
Small $\Delta = 0.05$, large $m = 300$



- Expected close to min → most trajectories find the right arm and incur no further regret after $t = mk$
- Confidence intervals of μ_1 and μ_2 have almost no overlap

Confidence intervals of per-arm reward

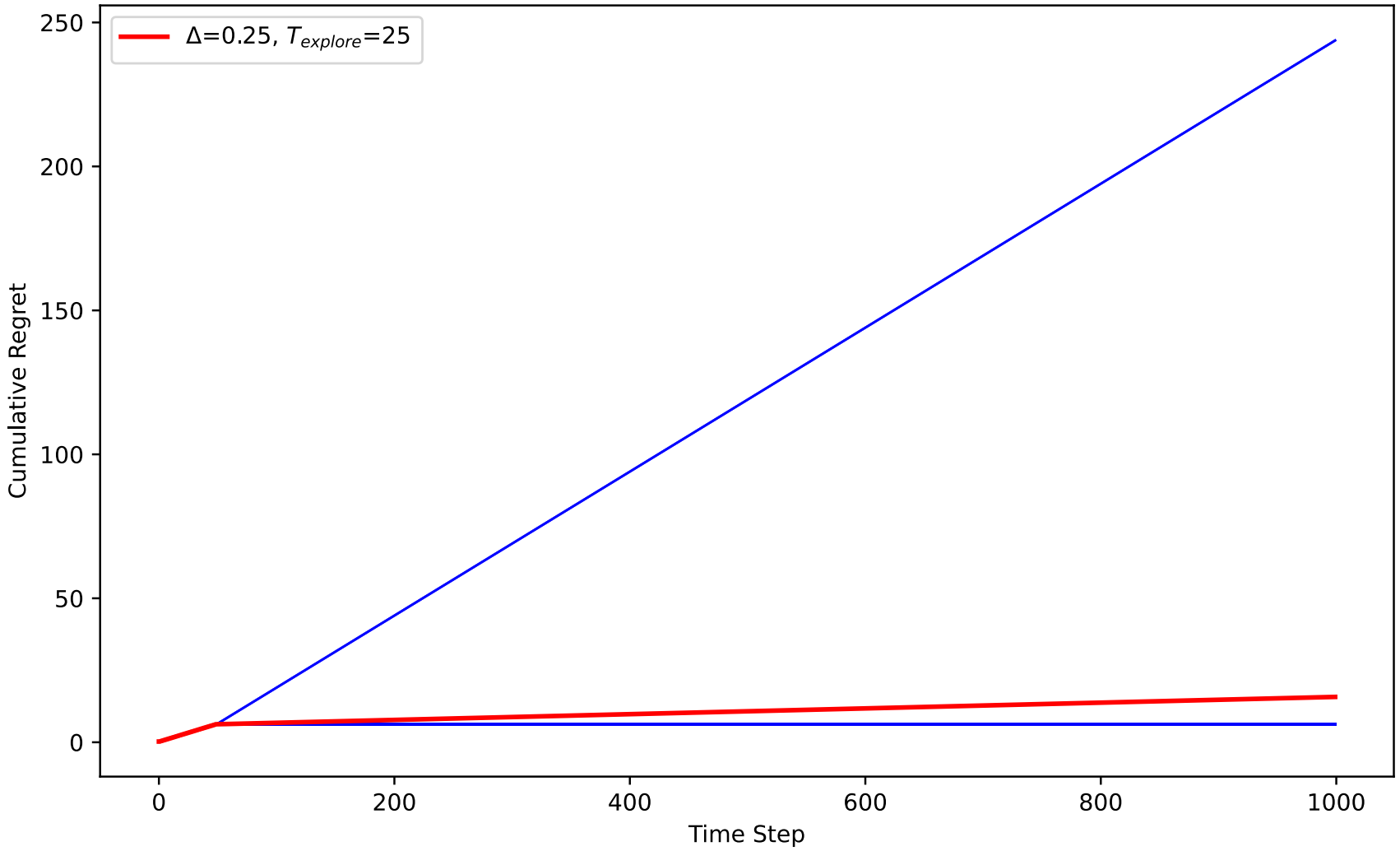
- Recall CI: $\left(\hat{\mu} - \sqrt{\frac{2\sigma^2 \ln(2/\delta)}{m}}, \hat{\mu} + \sqrt{\frac{2\sigma^2 \ln(2/\delta)}{m}}\right)$



- For $m = 25$:
- For $m = 100$:

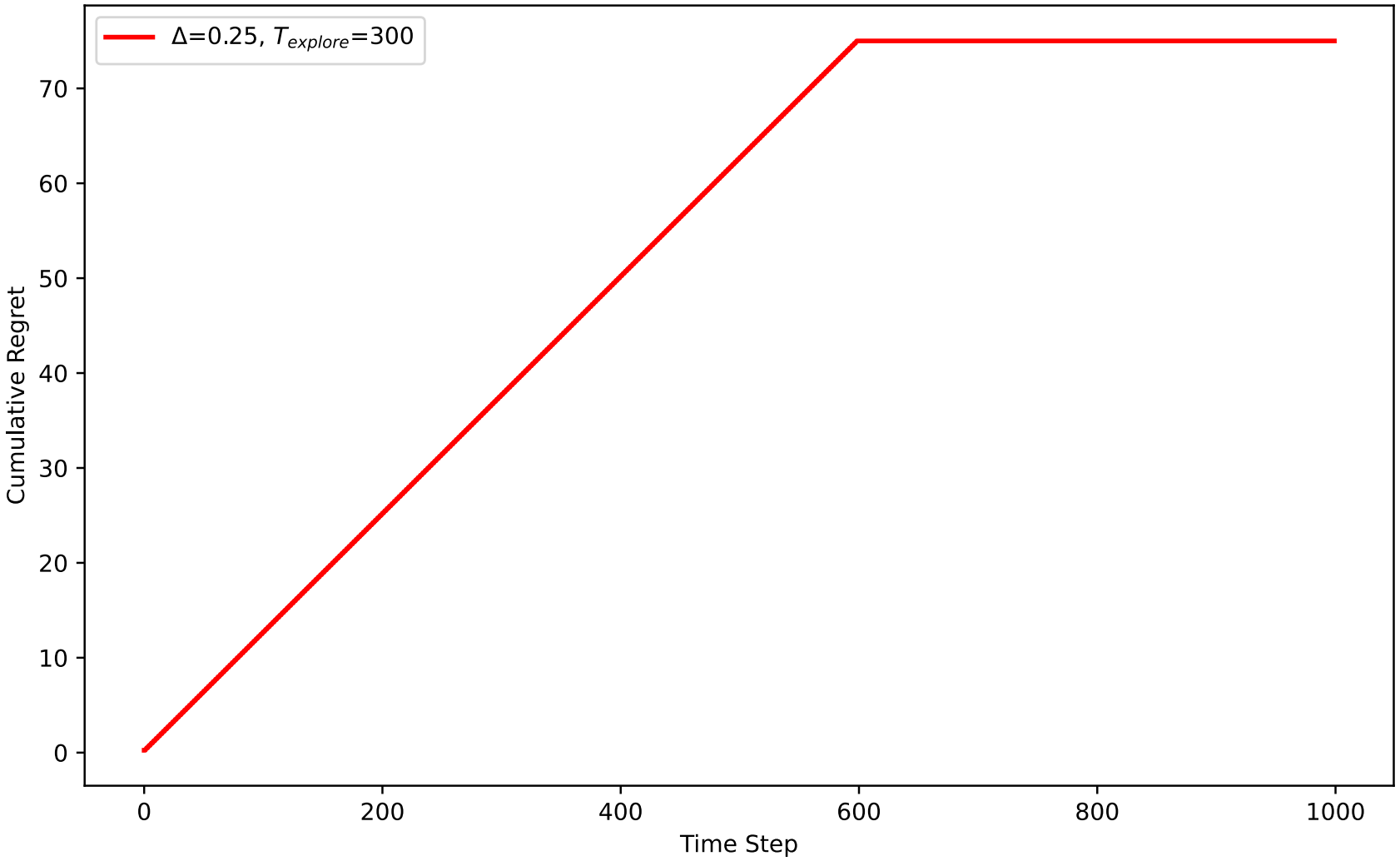
For $m = 300$:

Large $\Delta = 0.25$, small $m = 25$



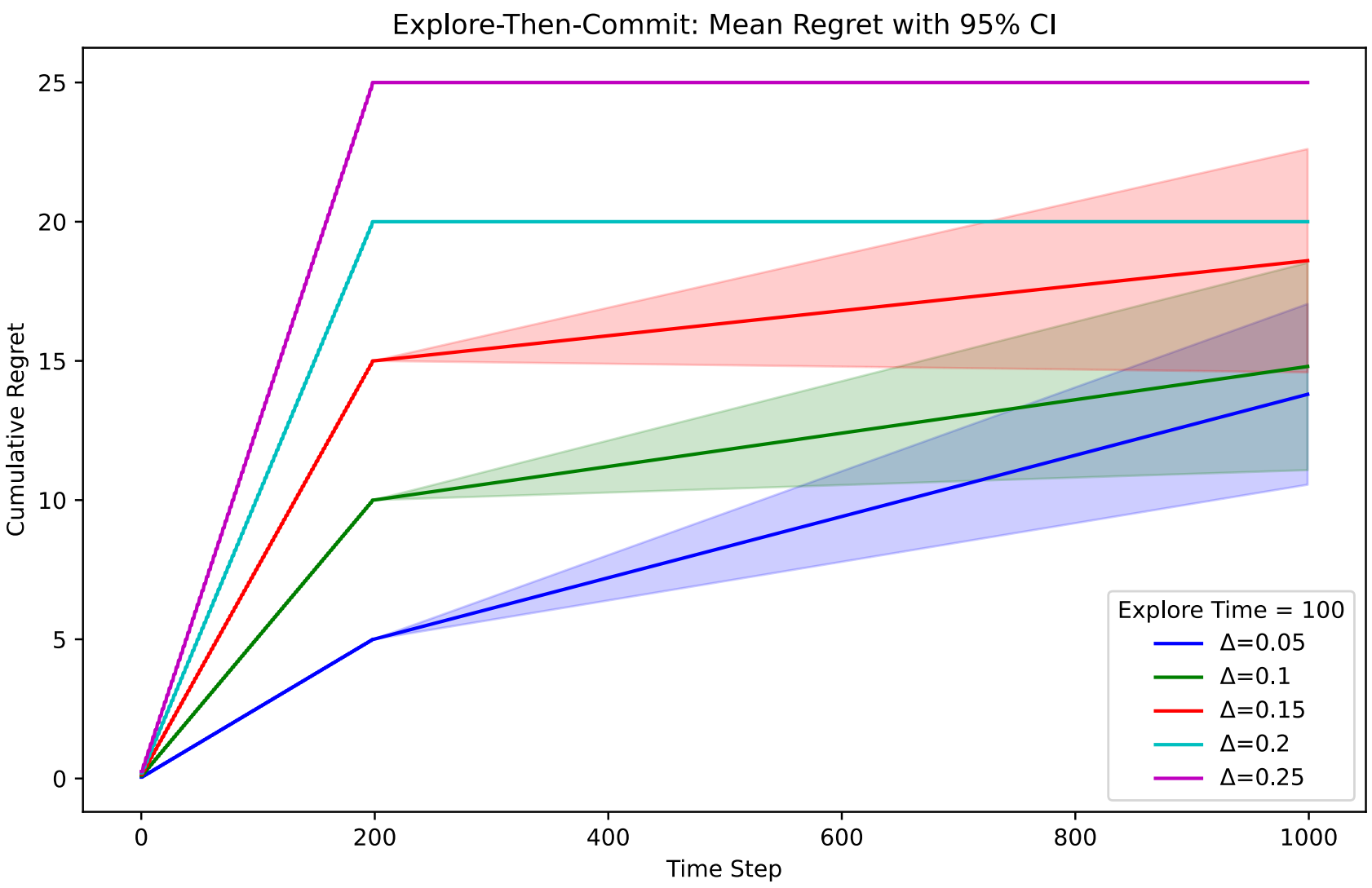
- Larger Δ means winning arm is more likely to be the better arm

Large $\Delta = 0.25$, large $m = 300$



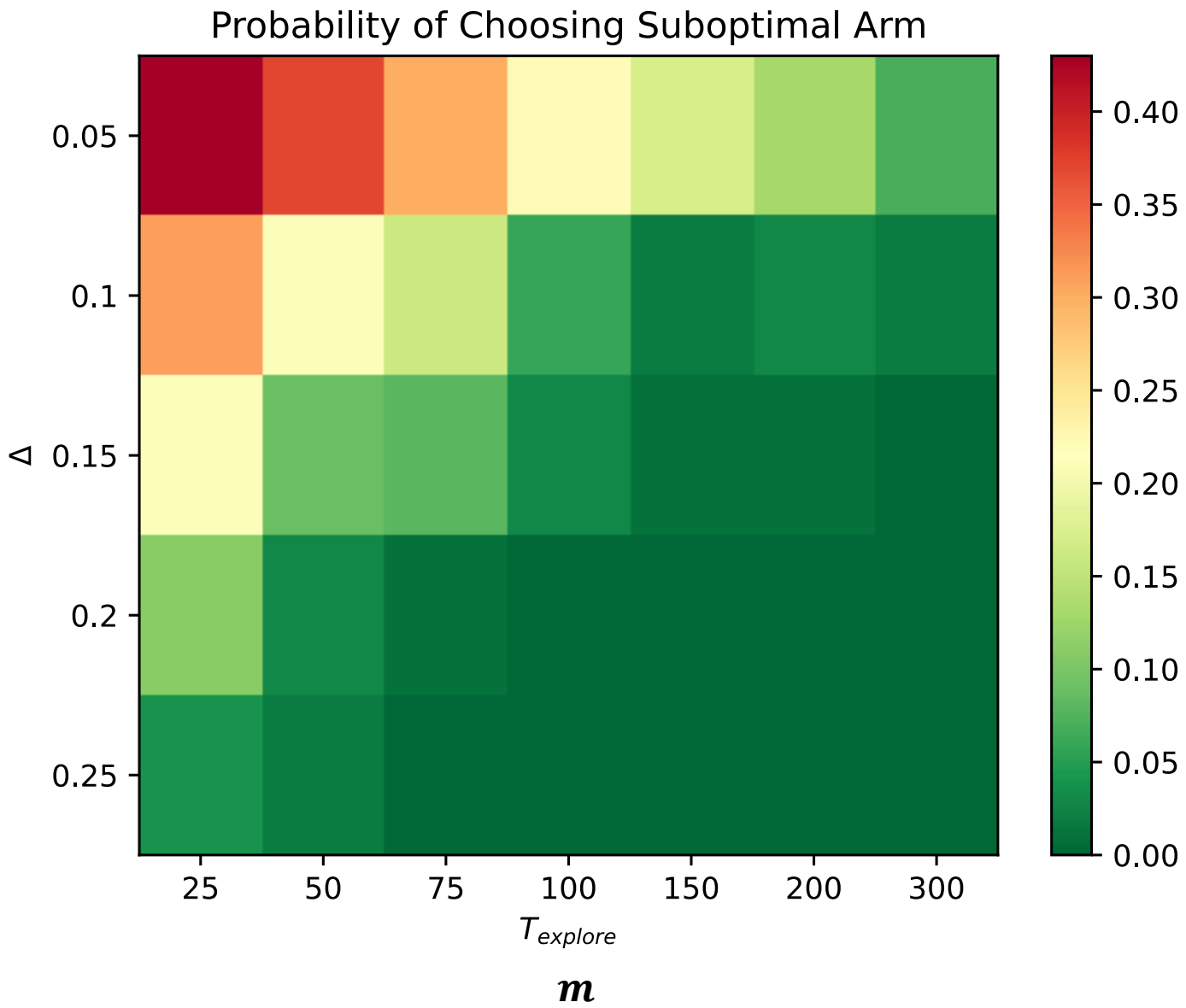
- No more failures in the simulation

Cumulative regret R_n for different Δ (for $m = 100$)



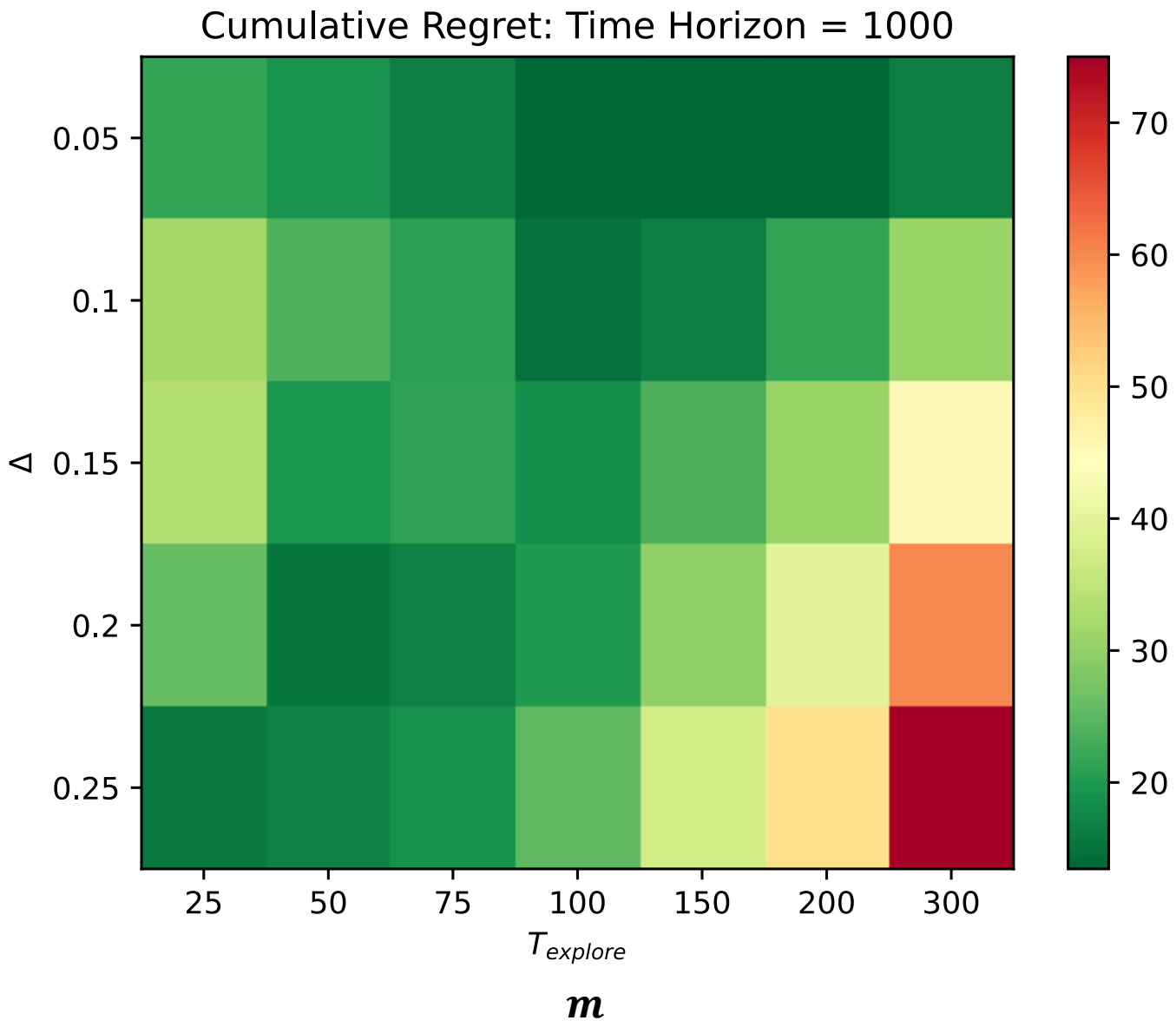
- Larger Δ means more costly exploration, but less risky exploitation

$\mathbb{P}(\text{wrong arm at time } n)$ as a function of Δ and k



- Increasing m helps finding the optimal arm
- This is made harder by smaller Δ

Cumulative regret R_n as a function of Δ and k



- Tradeoff in terms of cumulative regret:
  - Small Δ: needs large m
  - Large Δ: small m
  - Intermediate Δ: higher regret than the small/large Δ

Regret bound for 1-subgaussian bandits under ETC

- Thm: For ETC with the 1-subgaussian bandit,

$$R_n \leq \underbrace{m \sum_{i=1}^{k} \Delta_i}_{\text{exploration}} + \underbrace{(n - mk) \sum_{i=1}^{k} \Delta_i \exp\left(-\frac{m\Delta_i^2}{4}\right)}_{\text{exploitation}}$$

- Proof:
  - Assume (wlog) that $\mu_1 = \mu^* = \max_i \mu_i$
  - By decomposition lemma (cf last week),

$$R_n = \sum_{i=1}^{k} \Delta_i \mathbb{E}[T_i(n)]$$

  - $$\mathbb{E}[T_i(n)] = m + (n - mk)\mathbb{P}[A_{mk+1} = i]$$

$$\leq m + (n - mk)\mathbb{P}\left[\hat{\mu}_i(mk) \geq \max_{j \neq i} \hat{\mu}_j(mk)\right]$$

Regret bound for 1-subgaussian bandits under ETC

- Proof (cont.)
  - $\mathbb{P}\left[\hat{\mu}_i(mk) \geq \max_{j \neq i} \hat{\mu}_j(mk)\right] \leq \mathbb{P}[\hat{\mu}_i(mk) \geq \hat{\mu}_1(mk)]$
    $= \mathbb{P}[\hat{\mu}_i(mk) - \mu_i - (\hat{\mu}_1(mk) - \mu_1) \geq \Delta_i]$
    - Note: $m\hat{\mu}_i(mk)$ is $\sqrt{m}$-subgaussian, because it is the sum of m iid $\mathbf{1}$-subgaussians
    - Therefore, $\hat{\mu}_i(mk) - \hat{\mu}_1(mk)$ is $\sqrt{2/m}$-subgaussian
  - $\mathbb{P}[\hat{\mu}_i(mk) - \mu_i - (\hat{\mu}_1(mk) - \mu_1) \geq \Delta_i] \leq \exp\left(-\frac{m\Delta_i^2}{4}\right)$
- Tradeoff:
  - m large → exploration costly
  - m small → potential to bet on wrong arm forever

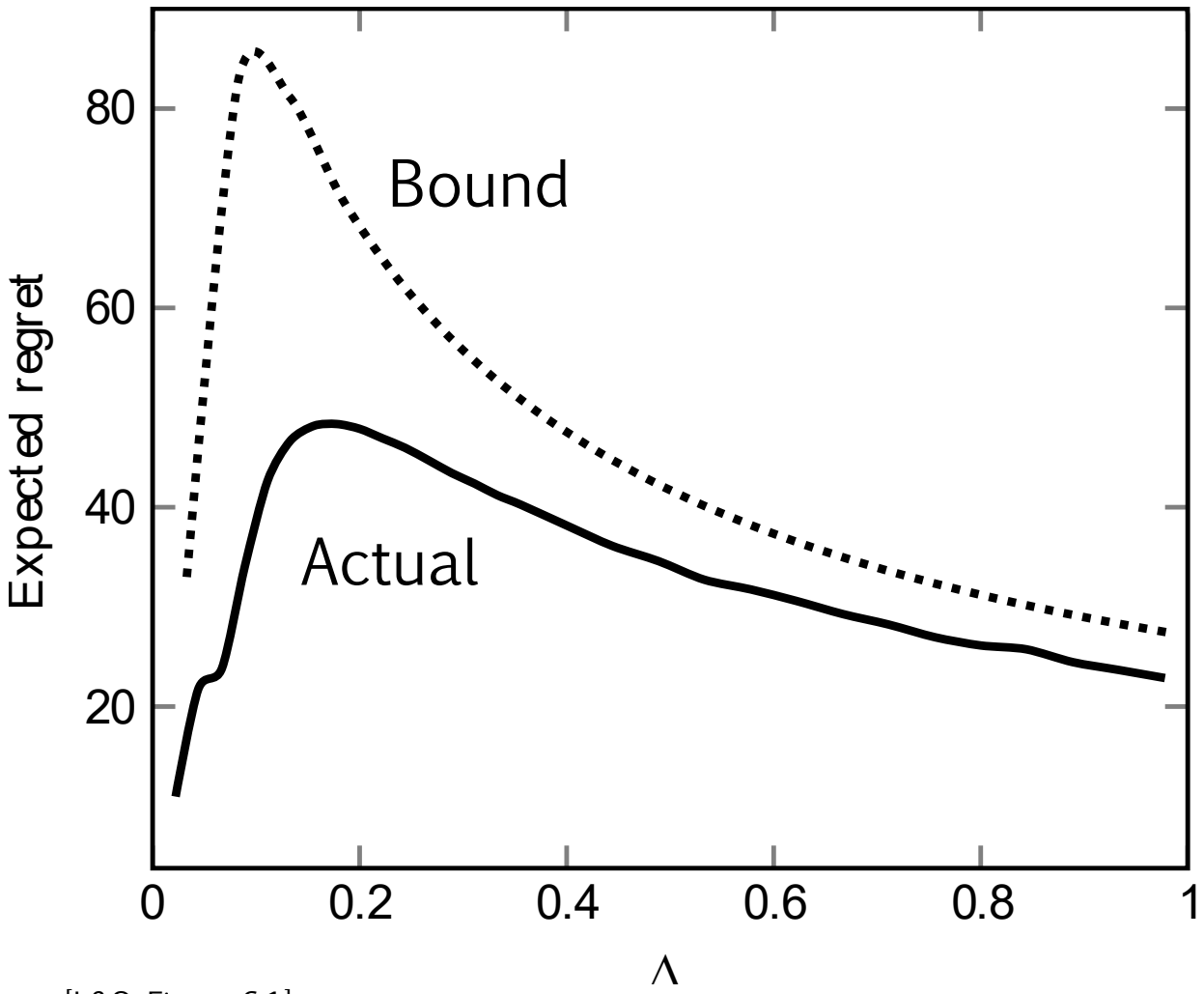
Analysis for $k = 2$

- Assume again that first arm is optimal: $\Delta_1 = 0$ (and call $\Delta_2 = \Delta$)
- Rewrite expression for general k:

$$R_n \leq m\Delta + (n - 2m)\Delta \exp\left(-\frac{m\Delta^2}{4}\right) \leq m\Delta + n\Delta \exp\left(-\frac{m\Delta^2}{4}\right)$$

- Which m minimizes this quantity?
  $m = \frac{4}{\Delta^2} \ln\left(\frac{n\Delta^2}{4}\right)$ (ignoring rounding)
- Resulting regret bound: $R_n \leq \Delta + C\sqrt{n}$
  - Without an assumption on Δ, cannot avoid potentially infinite regret (single bad pull)
  - Within bandit class such that $\Delta \leq 1$: $R_n \leq 1 + C\sqrt{n}$
- Example of a worst-case (or problem-independent) bound: does not depend on the specific environment instance ν
  - Only depends on environment class (here, subgaussian and $\Delta \leq 1$) and on horizon n
- But: to set the correct m, we **do** need to know Δ "ahead of time"
  - Can we have a universal algorithm that does well for unknown Δ?
  - Decide adaptively when we have explored enough?

Comparison of the bound with simulations



[L&S, Figure 6.1]

- The bound is relatively loose for intermediate subopt gap, but pretty good for larger
- Bound slightly underestimates the critical Δ

Summary

- What have we learned:
  - Convergence of random variables:
    - Sum of i.i.d.s: law of large numbers and central limit theorem
    - Tail bound for finite n: large deviation (Chernoff) bound
  - Subgaussian distributions: “nice” noise, light tail, “no big excesses”
    - Strong condition, even exponential is too much
    - Properties
  - Explore-then-Commit (ETC) algorithm:
    - Idea: explore for a fixed window, then exploit forever
    - Tradeoff between expensive exploration and failed exploitation
    - Upper bound for R_n for the 1-subgaussian bandit

- Reading assignment:
  - L&S: chapter 5, chapter 6