Lecture 2: Best-Arm Identification in Bandits

ID 6002W: Online and Reinforcement Learning

Web-enabled M.Tech. program, IIT Madras

Suryanarayana Sankagiri

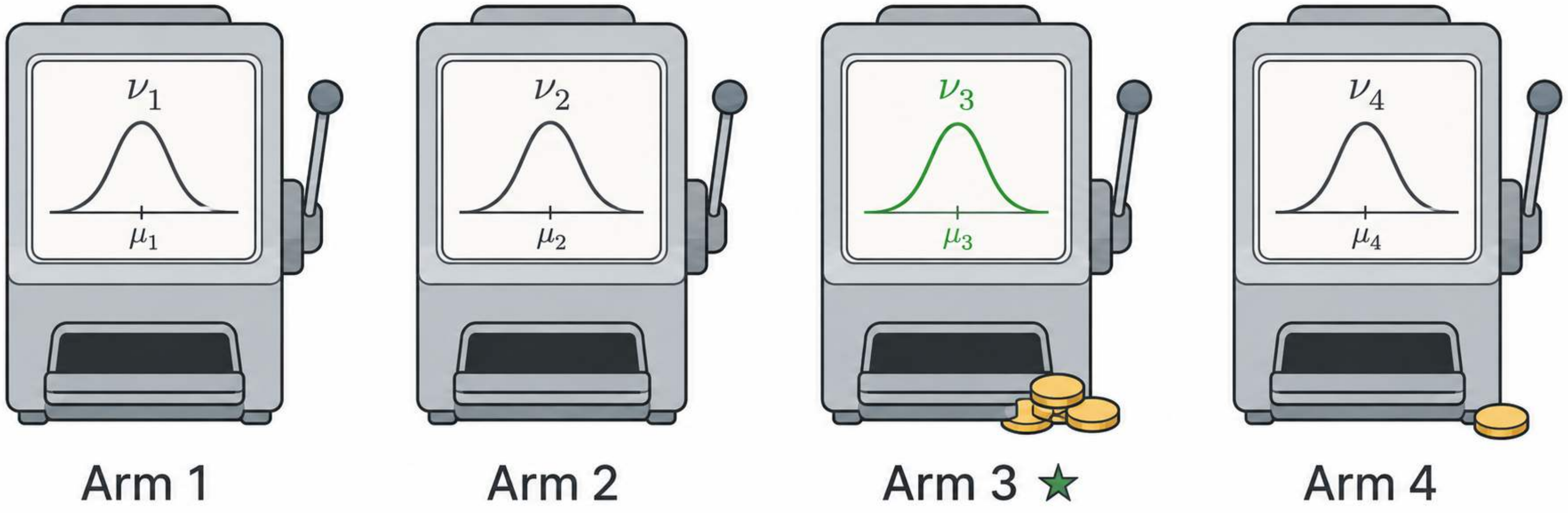
13th May, 2026

Where We Are:
Multi-Armed Bandits

Stochastic Multi-Armed Bandit

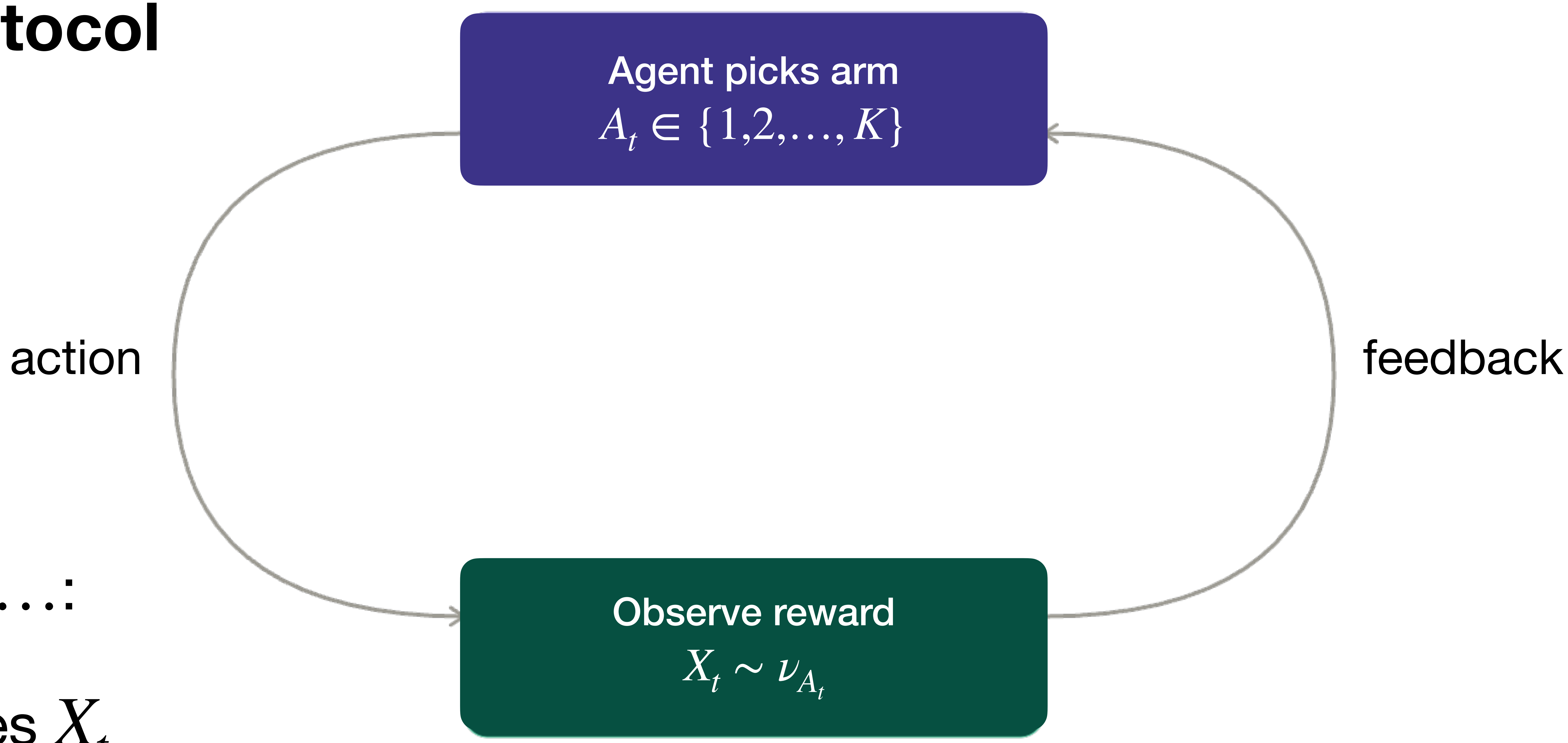
Basic Setup

- K arms; each arm corresponds to a unique action one can take
- arm i has reward distribution ν_i with mean μ_i
- Best arm: $i^* = \arg\max_{i \in \{1,2,\ldots,K\}} \mu_i$



Stochastic Multi-Armed Bandit

The interaction protocol



Repeat for $t = 1,2,3,\ldots$

- At each round $t = 1,2,\ldots$:
- agent picks A_t, observes X_t
- Bandit feedback: rewards of un-pulled arms are not seen

Stochastic Multi-armed Bandit

Special Cases

Usually, we will assume all arms have distribution of *same type*

1. Binary rewards (0/1)

- $\nu_i = \text{Ber}(\mu_i)$
- Bernoulli distribution with *parameter* μ:

$$X = \begin{cases} 1 & \text{with probability } \mu \\ 0 & \text{with probability } 1 - \mu \end{cases}$$

- Expected value of this distribution:

$$\mathbb{E}[X] = \sum_{x \in \text{Range}(X)} x\mathbb{P}(X = x) = 0\mathbb{P}(X = 0) + 1\mathbb{P}(X = 1) = \mu$$

Example: Observe if user clicks or does not click on article

Click-through rate of article i = average fraction of people clicking article i (when shown) = μ_i

Best article (arm): one with highest click-through rate

Stochastic Multi-armed Bandit

Special Cases

Usually, we will assume all arms have distribution of *same type*

2. Real-valued rewards

- $\nu_i = \mathcal{N}(\mu_i, 1)$ (normal/Gaussian distribution)
- Gaussian distribution with *mean μ and variance σ^2*:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- We will assume, for simplicity, that variance of all arms is 1
- Gaussian distribution is assumed even in scenarios where we know reward is ≥ 0

Example: Observe how long a user stays on a news article

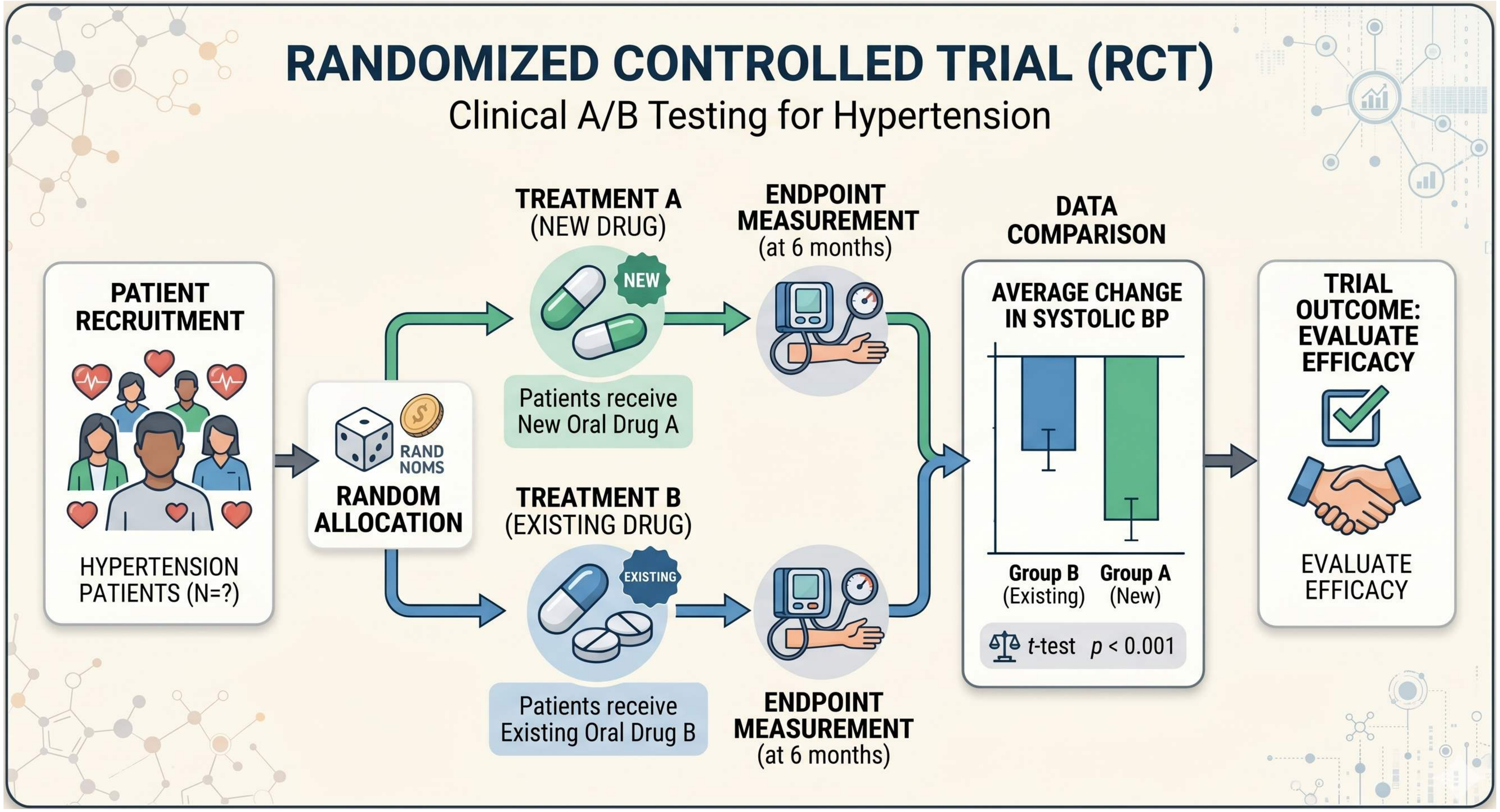
Average time spent on article i = μ_i

Best article (arm): one with highest engagement time

Today's Focus:
Best-Arm Identification

Best-Arm Identification

Example 1: Clinical Trials



Objective:

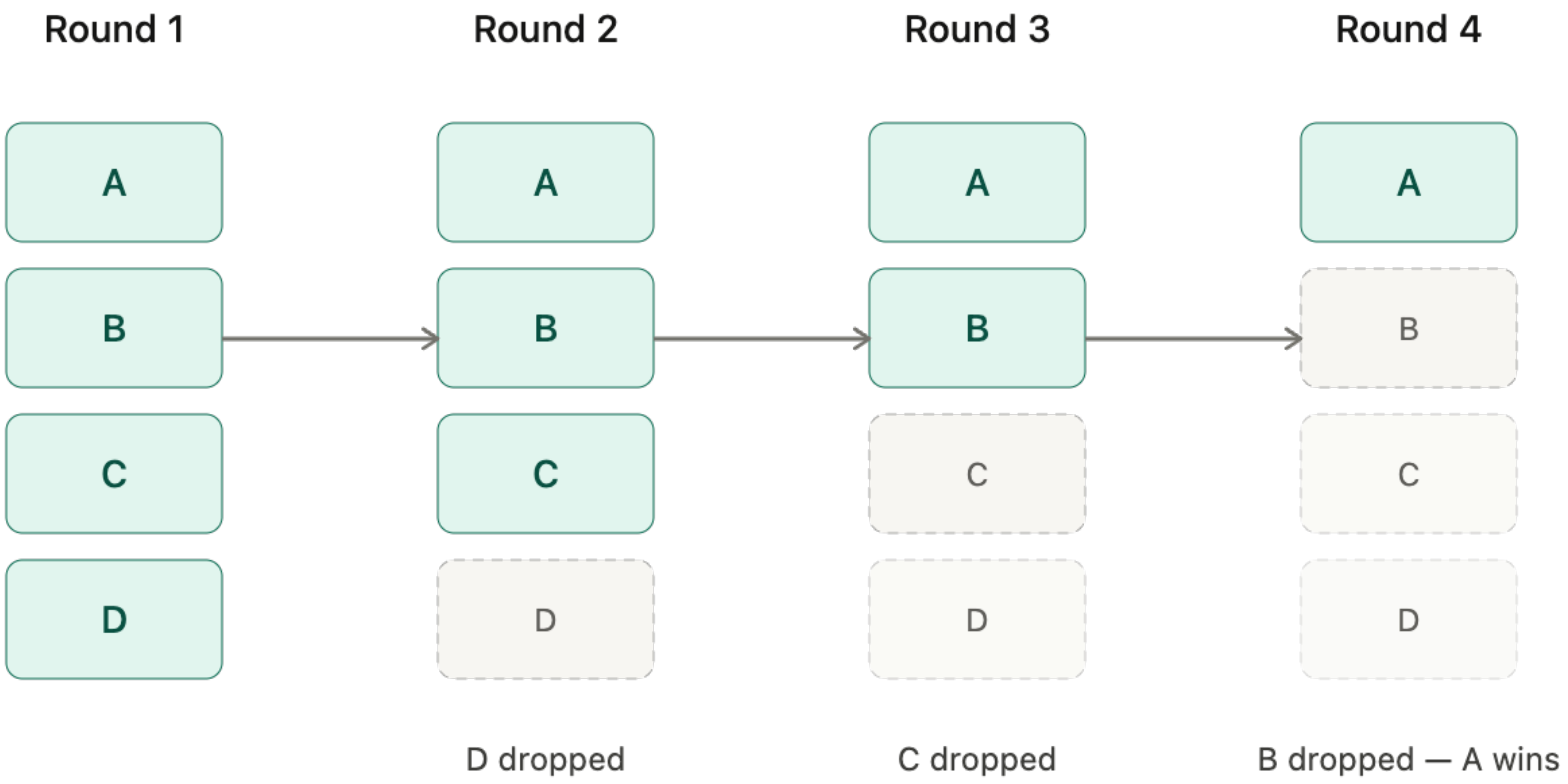
- Find best treatment with high confidence, using as few patients as possible
- Bad pulls *during* the trial are tolerated—cost of doing the trial properly

We call this the best-arm identification problem

Best-Arm Identification with Multiple Options

Example 2: New Feature Rollout

- **Setting:** Want to try out four new designs/layouts for a website, to find out the best one
- **Approach:** show different layouts to different users
- **Strategy:** Eliminate the clearly bad options as quickly as possible



Key decision at each step: Which arm to eliminate, and when?

Our Approach to Best-Arm Identification

How to think about today's lecture

- Goal: build intuition for stochastic bandits through a clean, simple problem
- Key tools we will use

  <u>confidence intervals</u> and <u>concentration inequalities</u>

  Will return for regret minimisation in the following lectures
- Best-arm identification (BAI) is a rich field; we'll only scratch the surface
  - Full treatment will require many weeks to cover properly (maybe even a PhD!)

Estimating Averages With Finite Samples

Estimating the Mean

Establishing Some Terminology

- Let X be a random variable, with distribution ν
- Let its <u>mean</u> be μ. $\mathbb{E}[X] = \mu$
    - If discrete: $\sum_{x \in \mathcal{X}} x \cdot \nu_x = \mu \qquad \left(\nu_x = \mathbb{P}(X = x)\right)$
    - If continuous: $\int_{x \in \mathcal{X}} x\nu(x)dx = \mu$
- When distribution ν is unknown, we cannot calculate μ. We can only estimate it

Estimating the Mean

Establishing Some Terminology

- How do we estimate the mean of a random variable X with distribution ν?
  - We need samples (in other words, data)
  - Independent and identically distributed (i.i.d.) samples

Estimating the Mean

Establishing Some Terminology

- How do we estimate the mean of a random variable X with distribution ν?
  - We need samples (in other words, data)
  - Independent and identically distributed (i.i.d.) samples
- When we pull one *fixed* arm i multiple times, the rewards $X_1, X_2, \ldots, X_n$ are i.i.d. samples from distribution ν_i
  - If you change arms from time to time, we need to be careful!
  - Need K separate buckets (or streams) of data, one for each arm

Estimating the Mean

Empirical Average versus Mean

Given i.i.d. samples $X_1, X_2, \ldots, X_n$ from some distribution,

- We can calculate the <u>empirical average</u> $\hat{\mu}(n)$: $\dfrac{X_1 + X_2 + \ldots + X_n}{n} = \dfrac{\sum_{t=1}^{n} X_t}{n}$
- The empirical mean $\hat{\mu}(n)$ is an <u>estimate</u> of μ

Estimating the Mean

Empirical Average versus Mean

Given i.i.d. samples $X_1, X_2, \ldots, X_n$ from some distribution,

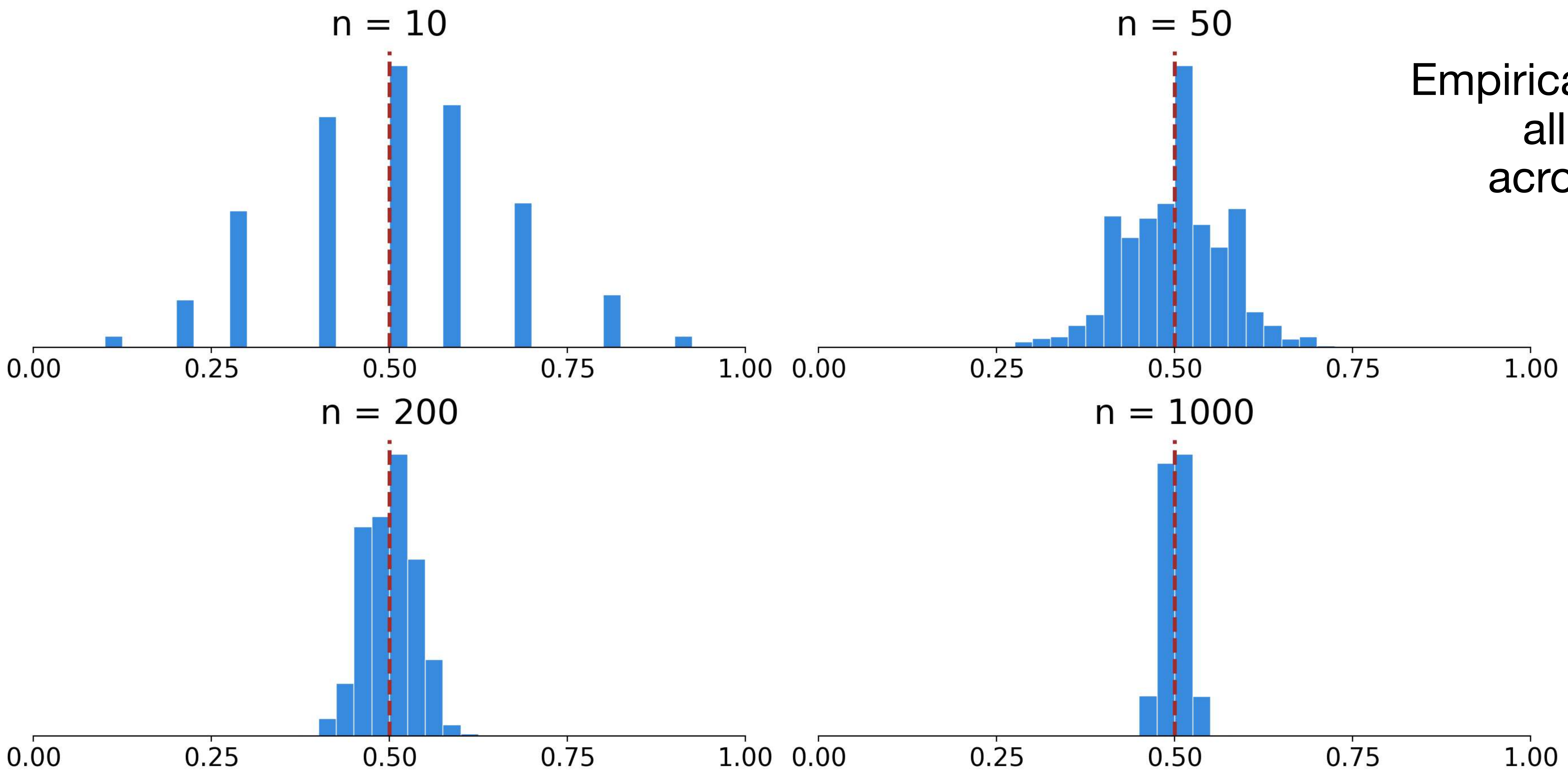
- We can calculate the empirical average $\hat{\mu}(n)$: $\dfrac{X_1 + X_2 + \ldots + X_n}{n} = \dfrac{\sum_{t=1}^{n} X_t}{n}$
- The empirical mean $\hat{\mu}(n)$ is an estimate of μ
- It may never actually be equal to μ

Law of Large Numbers says $\hat{\mu}(n) \rightarrow \mu$ as $n \rightarrow \infty$. Not useful for us.

We want something that tells us how $\hat{\mu}(n)$ concentrates around μ

Estimating the Mean

More Samples → Tighter Estimates



Empirical average of n samples,
all i.i.d. Bernoulli(0.5)
across 5000 simulations

We can quantify this tightening by Hoeffding's inequality

Hoeffding’s Inequality

Hoeffding's Inequality

Statement

Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in [a, b]$. Let $\mathbb{E}[X_t] = \mu$.

Let $\hat{\mu}(n) = \frac{1}{n}\sum_{t=1}^{n} X_t$. Then for any $\varepsilon > 0$,

Hoeffding's Inequality

Statement

Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in [a, b]$. Let $\mathbb{E}[X_t] = \mu$.

Let $\hat{\mu}(n) = \frac{1}{n} \sum_{t=1}^{n} X_t$. Then for any $\varepsilon > 0$,

$$\mathbb{P}(|\hat{\mu}(n) - \mu| \geq \epsilon) \leq 2 \exp\left(\frac{-2n\epsilon^2}{(b-a)^2}\right)$$

Hoeffding's Inequality

Statement

Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in [a, b]$. Let $\mathbb{E}[X_t] = \mu$.

Let $\hat{\mu}(n) = \frac{1}{n}\sum_{t=1}^{n} X_t$. Then for any $\varepsilon > 0$,

$$\mathbb{P}(|\hat{\mu}(n) - \mu| \geq \epsilon) \leq 2\exp\left(\frac{-2n\epsilon^2}{(b-a)^2}\right)$$

For Bernoulli case, $a = 0$, $b = 1$. So $\mathbb{P}(|\hat{\mu}(n) - \mu| \geq \epsilon) \leq 2\exp\left(-2n\epsilon^2\right)$

Hoeffding's Inequality

Interpreting The Bound

For Bernoulli case, $a = 0,\ b = 1$. So $\mathbb{P}(\,|\,\hat{\mu}(n) - \mu\,| \geq \epsilon) \leq 2\exp\left(-2n\epsilon^2\right)$

Multiple ways to interpret this formula.

Hoeffding's Inequality

Interpreting The Bound

For Bernoulli case, $a = 0,\ b = 1$. So $\mathbb{P}(\,|\,\hat{\mu}(n) - \mu\,| \geq \epsilon) \leq 2\exp\left(-2n\epsilon^2\right)$

Multiple ways to interpret this formula.

If you fix n and ϵ, we can calculate error probability δ

We have already seen: Probability of error shrinks *exponentially* in the number of samples

Hoeffding's Inequality

For Sample Complexity

For Bernoulli case, $a = 0,\ b = 1$. So $\mathbb{P}(|\hat{\mu}(n) - \mu| \geq \epsilon) \leq 2\exp\left(-2n\epsilon^2\right)$

Another interesting use of this equation is for calculating sample complexity

- How many samples (i.i.d. data points) do we need to solve the learning problem? Gives theoretical guarantees

Hoeffding's Inequality

For Sample Complexity

For Bernoulli case, $a = 0,\ b = 1$. So $\mathbb{P}(|\hat{\mu}(n) - \mu| \geq \epsilon) \leq 2\exp\left(-2n\epsilon^2\right)$

Another interesting use of this equation is for calculating <u>sample complexity</u>

- How many samples (i.i.d. data points) do we need to solve the learning problem? Gives theoretical guarantees
- fix ϵ and the failure probability δ, solve for n

$$2\exp(-2n\epsilon^2) \leq \delta \Rightarrow n \geq \frac{1}{2\epsilon^2}\log\left(\frac{2}{\delta}\right)$$

Hoeffding's Inequality

For Confidence Intervals

For Bernoulli case, $a = 0,\ b = 1$. So $\mathbb{P}(\,|\,\hat{\mu}(n) - \mu\,|\, \geq \epsilon) \leq 2\exp\left(-2n\epsilon^2\right)$

Hoeffding's Inequality

For Confidence Intervals

$$2\exp(-2n\epsilon^2) \leq \delta \Rightarrow \epsilon \geq \sqrt{\frac{\log(2/\delta)}{2n}}$$

With probability $\geq 1 - \delta$:

$$\mu \in \left[\hat{\mu} - \sqrt{\frac{\log(2/\delta)}{2n}}, \hat{\mu} + \sqrt{\frac{\log(2/\delta)}{2n}}\right]$$

Hoeffding's Inequality

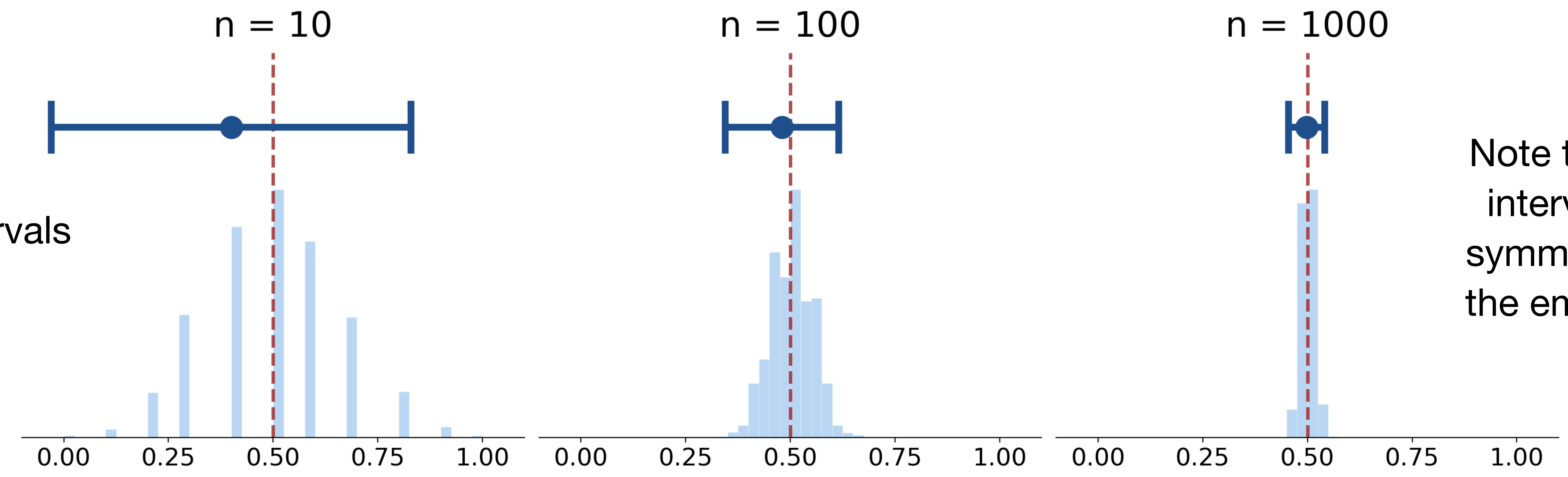
For Confidence Intervals

$$2\exp(-2n\epsilon^2) \leq \delta \Rightarrow \epsilon \geq \sqrt{\frac{\log(2/\delta)}{2n}}$$

With probability $\geq 1 - \delta$: $$\mu \in \left[\hat{\mu} - \sqrt{\frac{\log(2/\delta)}{2n}}, \hat{\mu} + \sqrt{\frac{\log(2/\delta)}{2n}}\right]$$

$95\ \%$ confidence intervals

$\delta = 0.05$



Note that confidence intervals are drawn symmetrically around the empirical average

Best-Arm Identification via Successive Elimination

Successive Elimination

The Key Idea

We have K arms. We want to find the best one with the following strategy.

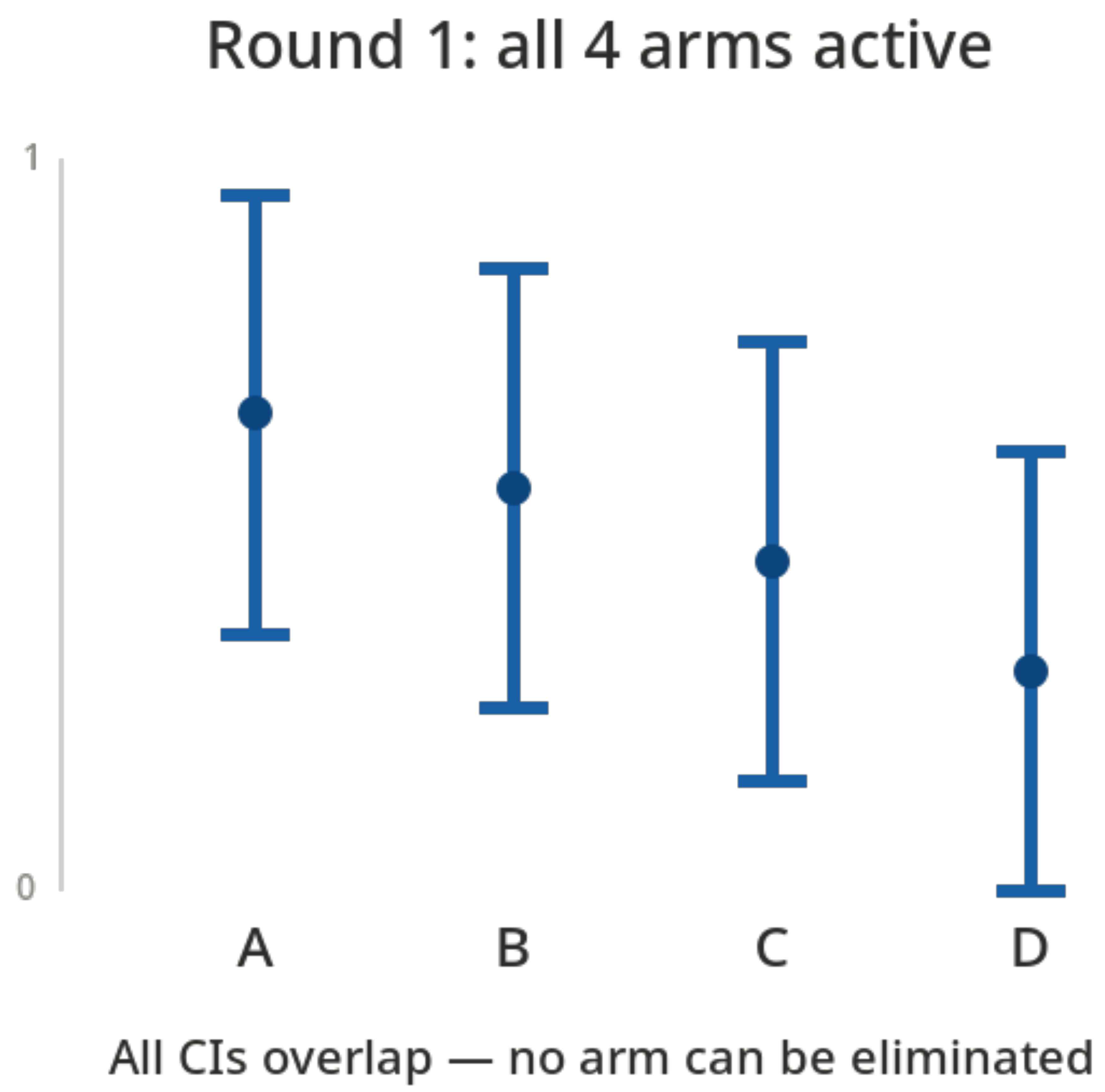
Successive Elimination

The Key Idea

We have K arms. We want to find the best one with the following strategy.

Maintain a set of **active arms** (initially, all K).
At each round t

- pull every active arm once
- Update *empirical means*, *confidence intervals*



Successive Elimination

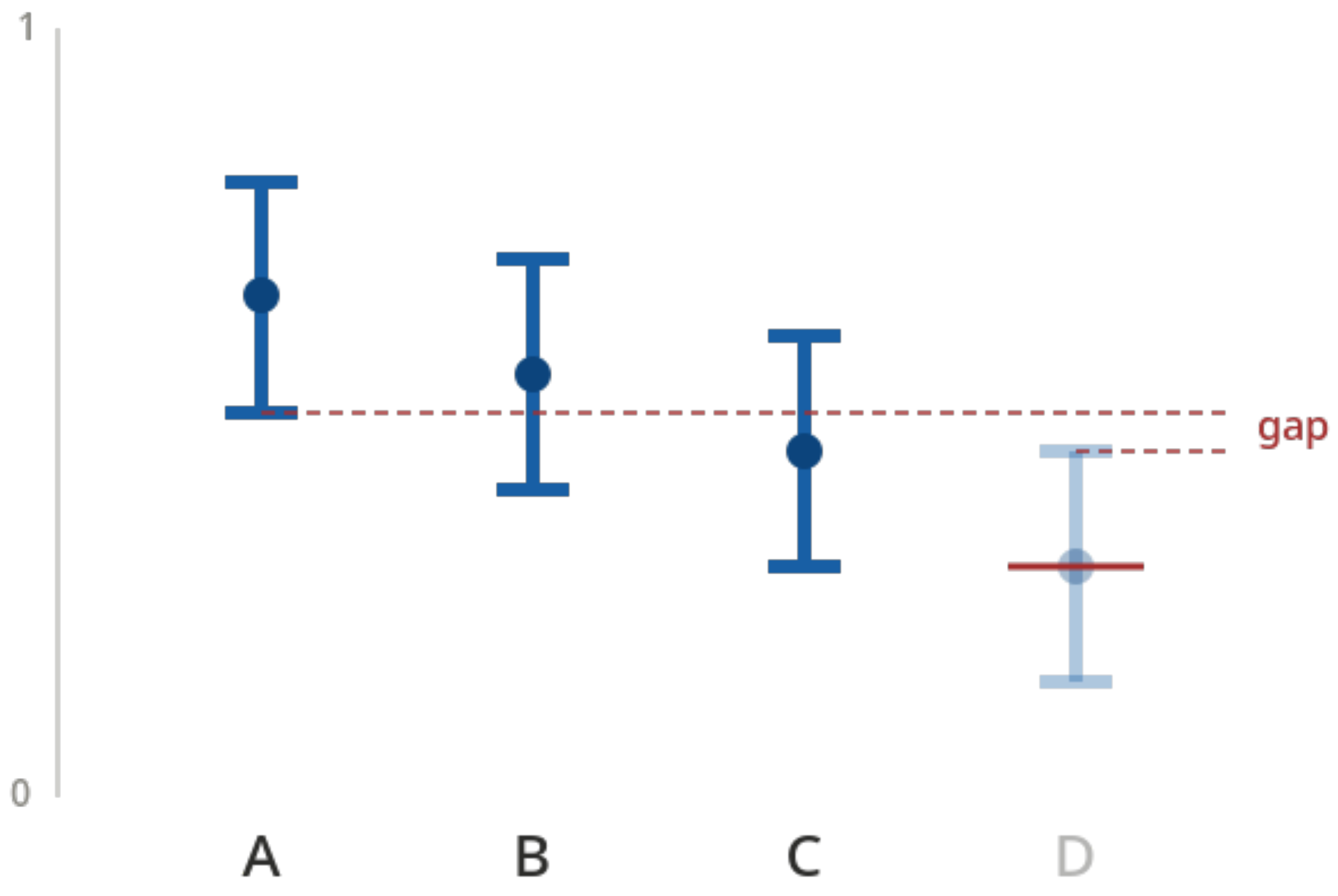
The Key Idea

We have K arms. We want to find the best one with the following strategy.

Maintain a set of **active arms** (initially, all K).
At each round t

- pull every active arm once
- Update *empirical means*, *confidence intervals*
- Eliminate any arm whose confidence interval lies *entirely below* another arm's interval

Round 5: D is eliminated



D's CI lies entirely below A's CI

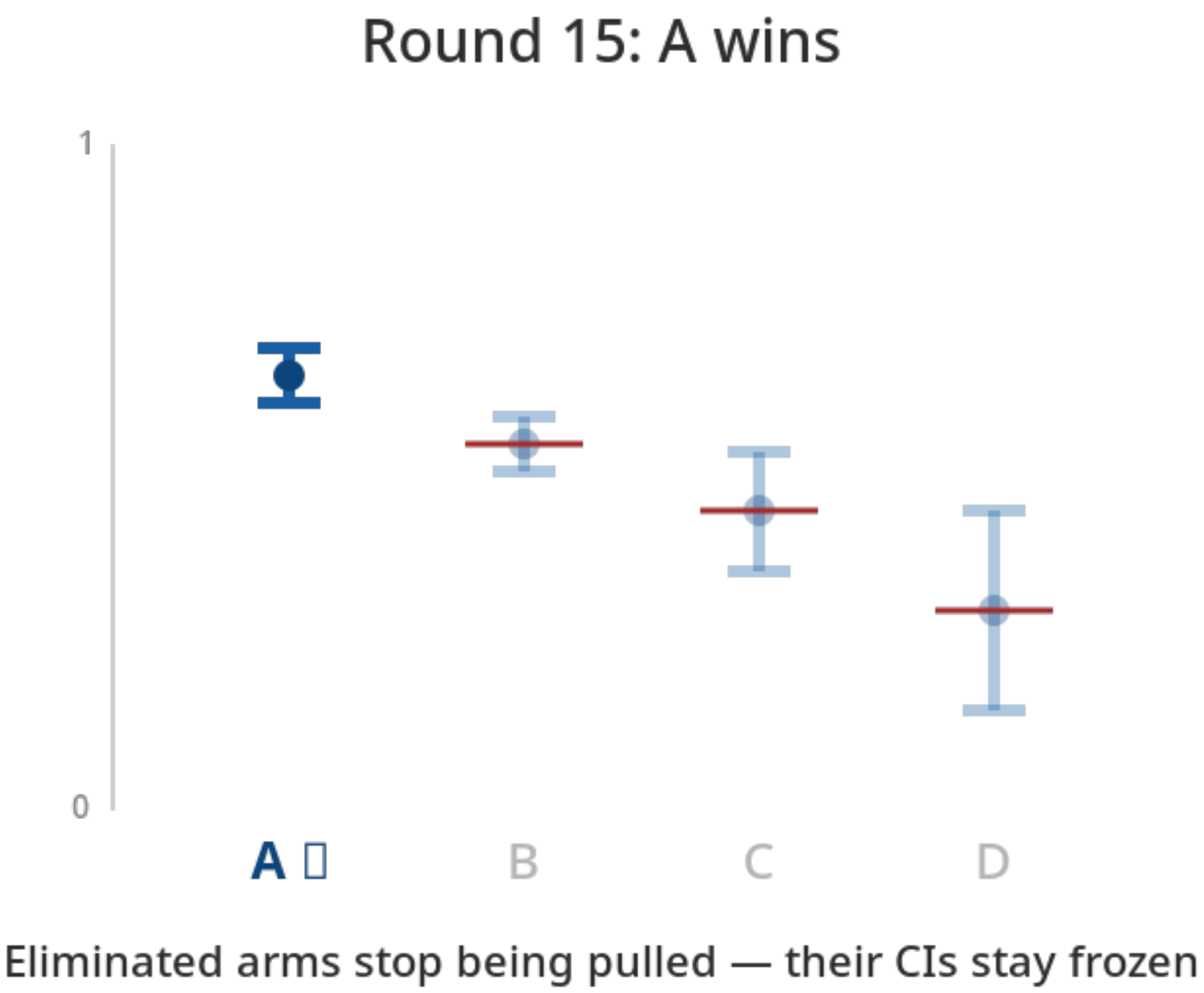
Successive Elimination

The Key Idea

We have K arms. We want to find the best one with the following strategy.

Maintain a set of **active arms** (initially, all K).
At each round t

- pull every active arm once
- Update *empirical means*, *confidence intervals*
- Eliminate any arm whose confidence interval lies *entirely below* another arm's interval
- Stop when only one arm remains



Successive Elimination

In Mathematical Notation

- $\text{UCB}_i(t) = \hat{\mu}_i(t) \ + \text{width}\,(t, \delta)$
- $\text{LCB}_i(t) = \hat{\mu}_i(t) \ - \text{width}\,(t, \delta)$

$$\text{width}_i\,(t, \delta) = \sqrt{\frac{\log(2/\delta)}{2t}}$$

Hoeffding's inequality applied at confidence $1 - \delta$

Elimination Rule

After round t, for any active arms i, j: if $\text{UCB}_i(t) < \text{LCB}_j(t)$, eliminate arm i

Successive Elimination

In Mathematical Notation

- $\mathrm{UCB}_i(t) = \hat{\mu}_i(t) \ + \text{width}\,(t, \delta)$
- $\mathrm{LCB}_i(t) = \hat{\mu}_i(t) \ - \text{width}\,(t, \delta)$

$$\text{width}_i\,(t, \delta) = C \sqrt{\frac{\log(2/\delta)}{2t}}$$

Hoeffding's inequality applied at confidence $1 - \delta$

Elimination Rule

After round t, for any active arms i, j: if $\mathrm{UCB}_i(t) < \mathrm{LCB}_j(t)$, eliminate arm i

Why It Works

- With high probability, every interval contains its true mean *throughout the run*
- If true, we never eliminate the best arm

- True mean μ^* lies inside its interval
- Any eliminated arm i has $\mathrm{UCB}_i < LCB^* < \mu^*$

Successive Elimination

How Long Does It Take?

Successive elimination with confidence parameter δ identifies the best arm with probability $\geq 1 - \delta$, and with total number of pulls at most:

$$C \cdot \Sigma_{i \neq i^*} \left(\frac{1}{\Delta_i^2} \right) \log \left(\frac{K}{\delta} \right) \quad \text{(}C\text{ is a constant)}$$

where $\Delta_i = \mu^* - \mu_i$ is the **gap** between arm i and the best arm.

Successive Elimination

How Long Does It Take?

Successive elimination with confidence parameter δ identifies the best arm with probability $\geq 1 - \delta$, and with total number of pulls at most:

$$C \cdot \Sigma_{i \neq i*} \left(\frac{1}{\Delta_i^2} \right) \log \left(\frac{K}{\delta} \right) \quad \text{(}C\text{ is a constant)}$$

where $\Delta_i = \mu^* - \mu_i$ is the **gap** between arm i and the best arm.

Intuition: Arm i eliminated when its C.I. sits entirely below i^* C.I.
$\Leftrightarrow$ combined width of the two C.I.s is smaller than the gap Δ_i

Width of C.I. reduces proportional to $1/\sqrt{t}$

What About Regret Minimisation?

How does the intuition carry over?

- Next lecture, we will see the Explore-Then-Commit algorithm
  - Very similar to sequential elimination
    - Analysis via confidence intervals (obtained from Hoeffding's inequality)
  - But, we will be accounting for regret!

What About Regret Minimisation?

How does the intuition carry over?

- Next lecture, we will see the Explore-Then-Commit algorithm
  - Very similar to sequential elimination
    - Analysis via confidence intervals (obtained from Hoeffding's inequality)
  - But, we will be accounting for regret!

Which is the right approach?

- *BAI is right* when you'll commit to one decision after a finite search, and intermediate pulls are cheap (feature rollouts, hyperparameter tuning, offline experimentation)

- *Regret minimisation is right* when every action has real-world cost; learning and earning are the same activity (recommendations, dynamic pricing, ad serving).

Conclusion

What We Learned Today

- **Best-arm identification (BAI) in multi-armed bandits.**
    - A clean, natural objective in online learning:
    - Find the best action with high confidence, using as few samples as possible
- **Hoeffding's inequality and confidence intervals.**
    - They turn "I'm pretty confident" into a quantitative confidence interval
    - Core tool in this course and many aspects of data science
- **Successive elimination:** an algorithm for BAI that uses confidence intervals to drop arms adaptively