Module 7
Multi-armed bandits

DAV-6300-1: Experimental Optimization

David Sweet // 20241029

Review: Randomization

- A/B test: A=old ad, B=new ad
- Business metric is ad revenue/day
- A/B test design says N=10,000
- The A/B test has been running for three days, and you've collected 4,000 observations each of A and B so far. You calculate t from the 4,000 ind. meas:
    - $t = \frac{\mu}{se} = 8.3$ <== 8.3 is large. What does this tell you?

Review: Early Stopping

- $t = \frac{\mu}{se} = 8.3$ <== What does this tell you?
- Note: $se = \frac{\sigma_\delta}{\sqrt{4000}}$ > $se = \frac{\sigma_\delta}{\sqrt{10000}}$
- Therefore μ_B must be *much* larger than μ_A

Review: Early Stopping

- Stop now, capture extra revenue from B
  - I.e., reduce opportunity cost
- But, early stopping leads to false positives
- What could we do?

Key Terms

- Exploration
- Exploitation
- Arm
- Multi-armed bandit

Multi-armed bandits

Motivation

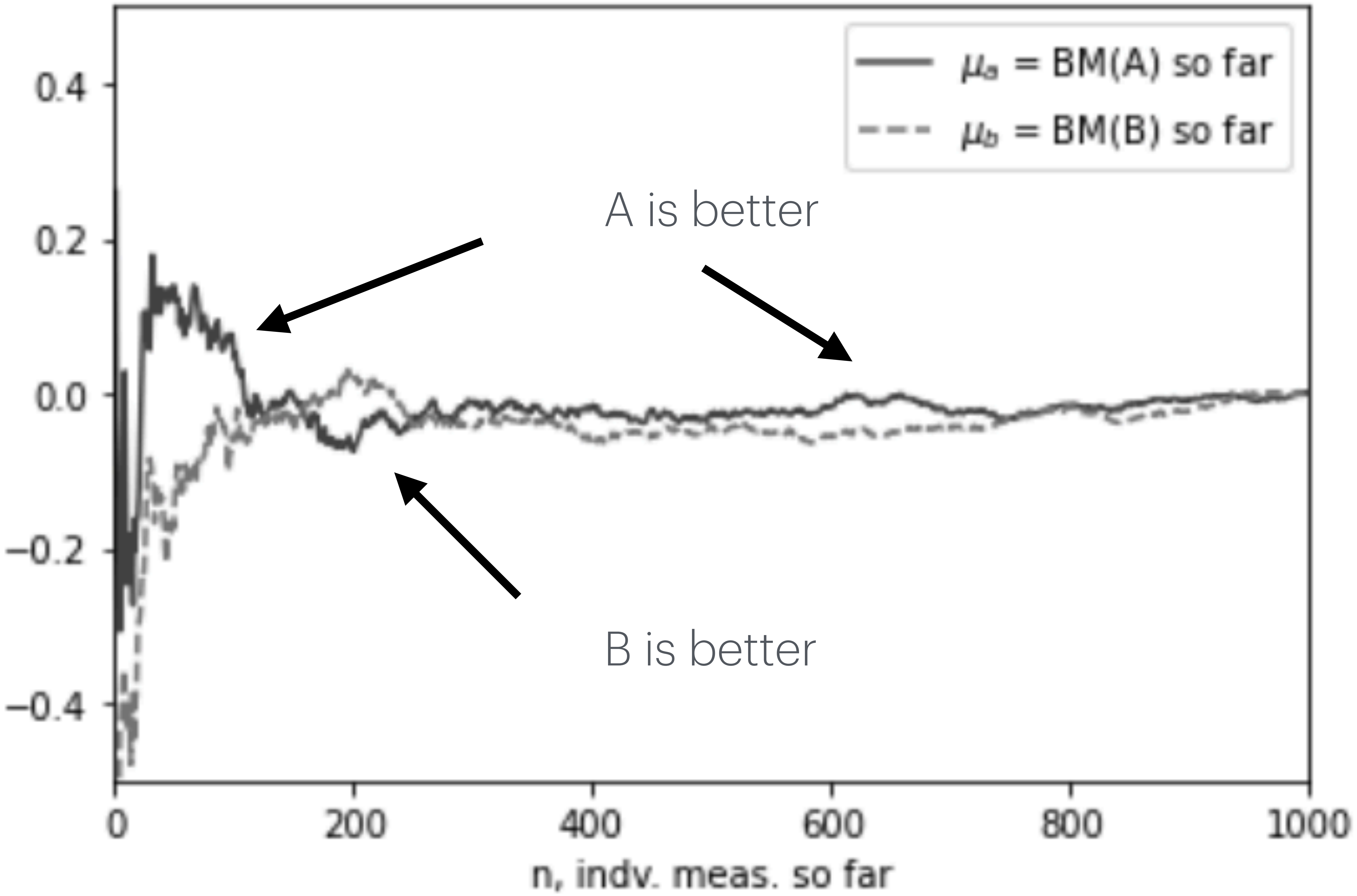
- Note 1: FP/FN errors are more common when BM(B) is closer in value to BM(A).
- Note 2: We’re interested in optimizing **business metric**, not FP/FN rates.
  - Want more revenue, more clicks, less fraud, etc.
- FP/FN rates tell the quality of the experiment.
  - BM tells the quality of the business.

Multi-armed bandits

This is early stopping

- Proposal I: At any point during the experiment, just run whichever version, A or B, has the higher BM.

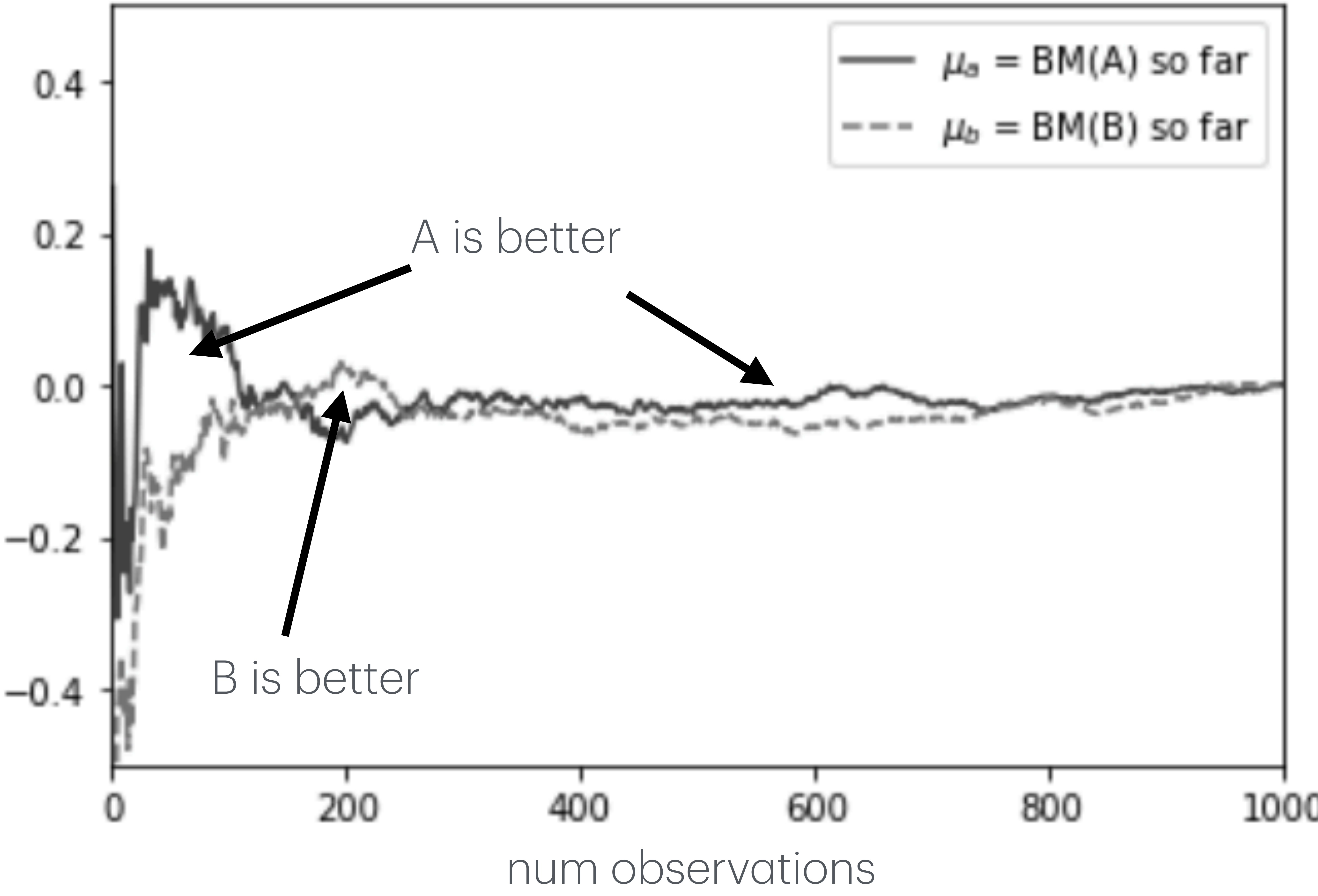
- Problem: Could accidentally turn off better version



Multi-armed bandits

Not stopping
No decision == > No FP

- Proposal II: **Usually** run whichever version, A or B, has the higher BM.
- "usually": Assign 90% of observations to better (so far) of A & B
- 10% of time, choose A,B randomly



Multi-armed bandits

- "10% of time, choose A,B randomly": keeps collecting observations of "worse" version
  - Allows BM estimate of worse version to continue to vary
    - Maybe later on this will be the better version
  - Reduces se's of both versions
  - Lower se's ==> more precise comparison

Multi-armed bandits

- How does this optimize the business metric?
- At any point during the experiment
  - Better BM-so-far ==> *probably* better expectation
  - 90% chance you're running with better expectation
  - Better overall BM while experimenting

Multi-armed bandits

Epsilon-greedy

- $\varepsilon = 0.10$ ("10% of the time")
- For every observation:
  - $p_{explore} = \varepsilon$: Choose A or B at random
  - $p_{exploit} = 1 - p_{explore} = 1 - \varepsilon$: Run version w/higher μ_n
- Exploitation helps you get higher BM **now**.
- Exploration improves BM estimates (reduces SE), so you get higher BM in the **future**.

"Balance exploration with exploitation"

Multi-armed bandits

Epsilon-greedy

- $\mu_{n,A}$ - Mean of all A observations taken so far
- $\mu_{n,B}$ - Mean of all B observations taken so far
- $P_n\{FP\}$ - Probability that
  - $\mu_{n,A} > \mu_{n,B}$ but $E[A] < E[B]$, or
  - $\mu_{n,B} > \mu_{n,A}$ but $E[B] < E[A]$

Multi-armed bandits

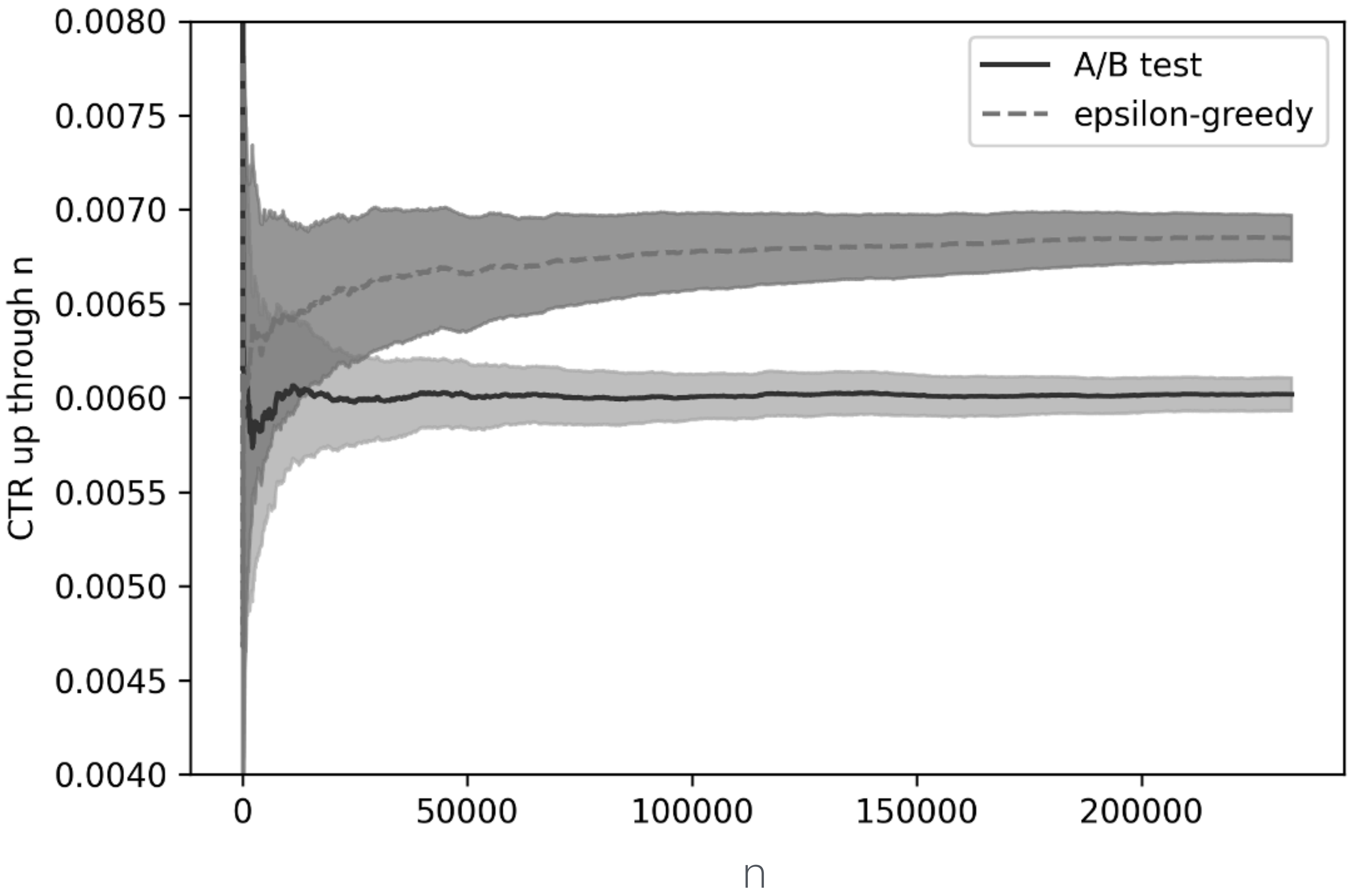
Epsilon-greedy

Probability of running the better version

	Better version	**Worse version**
Exploit	0.90 X (1 - P{FPn})	0.90 X P{FPn}
Explore	0.10 X 0.50	0.10 X 0.50

Multi-armed bandits

Epsilon-greedy



BM is higher during the experiment with ε-greedy

Multi-armed bandits

Epsilon-greedy summary

- **Maximize BM during experiment**: ε-greedy changes the goal of experiment design from “limit FP/FN” to “maximize BM while experimenting”

- **Usually run the better version (exploitation)**: ε-greedy modifies the randomization procedure of A/B testing from “50/50” to “90/10”. 90% of the time you run the version with higher BM-so-far.

- **Sometimes run the worse version (exploration)**: Exploration lowers SE of worse version to improve later decisions about which version is better. 10% of the time you run a version chosen at random.

Multi-armed bandits

Epsilon-greedy: When do you stop?

- There's no "N" in epsilon-greedy
- Could use N from A/B test design:
  - Find $N = \frac{\sqrt{N}\sigma_\delta}{PS}$
  - Run ε-greedy until both A and B have at least N observations
- How would the experimentation cost compare to an A/B test?

Multi-armed bandits

Epsilon-greedy: When do you stop?

- How would the experimentation cost compare to an A/B test?
    - You’d run the worse version N times
    - You’d run the better version more than N times b/c of the 90% rule
    - Thus, overall, this would take much longer to run than an A/B test
- You only “win” if you run the worse version fewer times than you would have in an A/B test, i.e., fewer than N times

Multi-armed bandits

Epsilon-greedy: When do you stop?

- Solution: Decrease ε over the course of the experiment.
- Start: $\varepsilon_0 = 0.1$
- On n^{th} observation: $\varepsilon_n \propto 1/n$
- Stop when ε_n is below some threshold, ex., $\varepsilon_{stop} = 0.01$, where exploration is insignificantly small.
- IOW, stop when not really experimenting any more

Multi-armed bandits

Epsilon-greedy: When do you stop?

- More precisely:

- $$\varepsilon_n = \frac{2c(BM_0/PS)^2}{n}$$

- BM_0 is a scale for your business metric

- PS is the same practical significance level from A/B test design

- $c = 5$

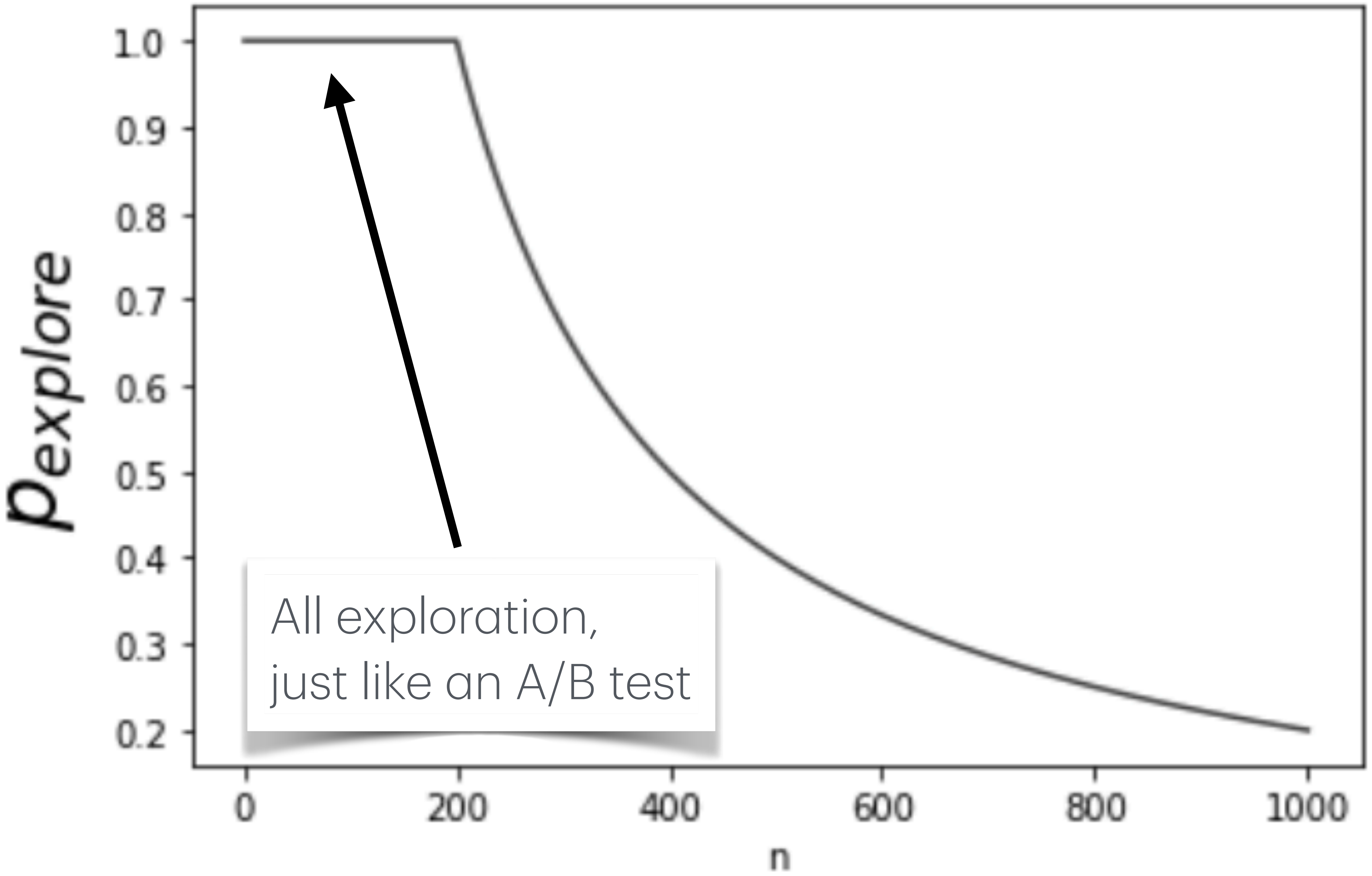
- Not pretty, but robust to your choices of BM_0, c, and ε_{stop}

Will a larger PS make this experiment run for more or less time?

Multi-armed bandits

Epsilon-greedy: When do you stop?

- Since probability can't be larger than one, practically speaking:
  - $p_{explore} = min(1, \varepsilon_n)$
  - $p_{exploit} = 1 - p_{explore}$
  - Optimal regret
    P. Auer, N. Cesa-Bianchi, and P. Fisher,
    "Finite-time analysis of the multiarmebandit problem,"
    Mach. Learn., vol. 47, 235–256, 2002



Multi-armed bandits

One more thing…

- In MAB lingo, A and B are called "arms" instead of versions.
- It's really easy to test more than two arms:
  - $p_{explore} = \varepsilon$: Run any arm — A, B, C, … — at random
  - $p_{exploit} = 1 - p_{explore} = 1 - \varepsilon$: Run the highest-BM-so-far of A, B, C, …
- IOW, usually run the best arm.

Multi-armed bandits

One more thing...

- Also, change this:
  - $\varepsilon_n = \dfrac{\mathbf{2}c(BM_0/PS)^2}{n}$

k=2, here, just A and B

- to this:
  - $\varepsilon_n = \dfrac{\mathbf{k}c(BM_0/PS)^2}{n}$

Sometimes called “k-armed bandit”

- where k is the number of arms.

Multi-armed bandits

Summary

- MAB goal: Maximize BM during the experiment, i.e. minimize experimentation cost
- Epsilon-greedy:
  - Exploit: Usually run the best arm
  - Explore: Sometimes run a random arm
  - Decay: Explore less as se's shrink
  - Stop: When exploration rate is tiny (i.e., not really experimenting any more)

Multi-armed bandits

Summary

- Easy to set up
- Easy to compare multiple arms
- Decay schedule is a little clunky

Final Project

- Two groups
- Optimize a black-box function
- One measurement / day
- See Final.ipynb
- http://cogneato.xyz