Experimental optimization

Lecture 7: Multi-armed bandits I: Epsilon-greedy

David Sweet

Review

Early stopping

- A/B test: A=old ad, B=new ad created by a new ad creator company
- 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 individual measurements each of A and B so far. You calculate z from the 4,000 ind. meas:
  - $z = \frac{\mu}{SE} = 8.3$ <== 8.3 is large. What does this tell you?

Review

Early stopping

- $z = \frac{\mu}{SE} = 8.3$ <== What does this tell you?
- Note: $SE = \frac{\sigma_\delta}{\sqrt{4000}}$
- Assuming that σ_δ is similar to your estimate from design time, for z to be large, it must be that μ is large, and
  - $\mu = \mu(B) - \mu(A) \sim BM(B) - BM(A)$
- Therefore B must be *much* better than A!

Review

Early stopping

- If B is much better than A, then you want to stop the A/B test and switch over to B to capture the extra revenue.

- If you stop early b/c z is large, what bad thing happens?

Review

Early stopping

- If B is much better than A, then you want to stop the A/B test and switch over to B.
- If you stop early b/c z is large, what bad thing happens?
  - You increase the risk of a false positive (by a lot!)
- But, you run lots of experiments, and you worry that waiting a few more days for experiments to complete when the result seems *obvious* is just a waste of money, time, etc. — **experimentation costs**.

Multi-armed bandits

Motivation

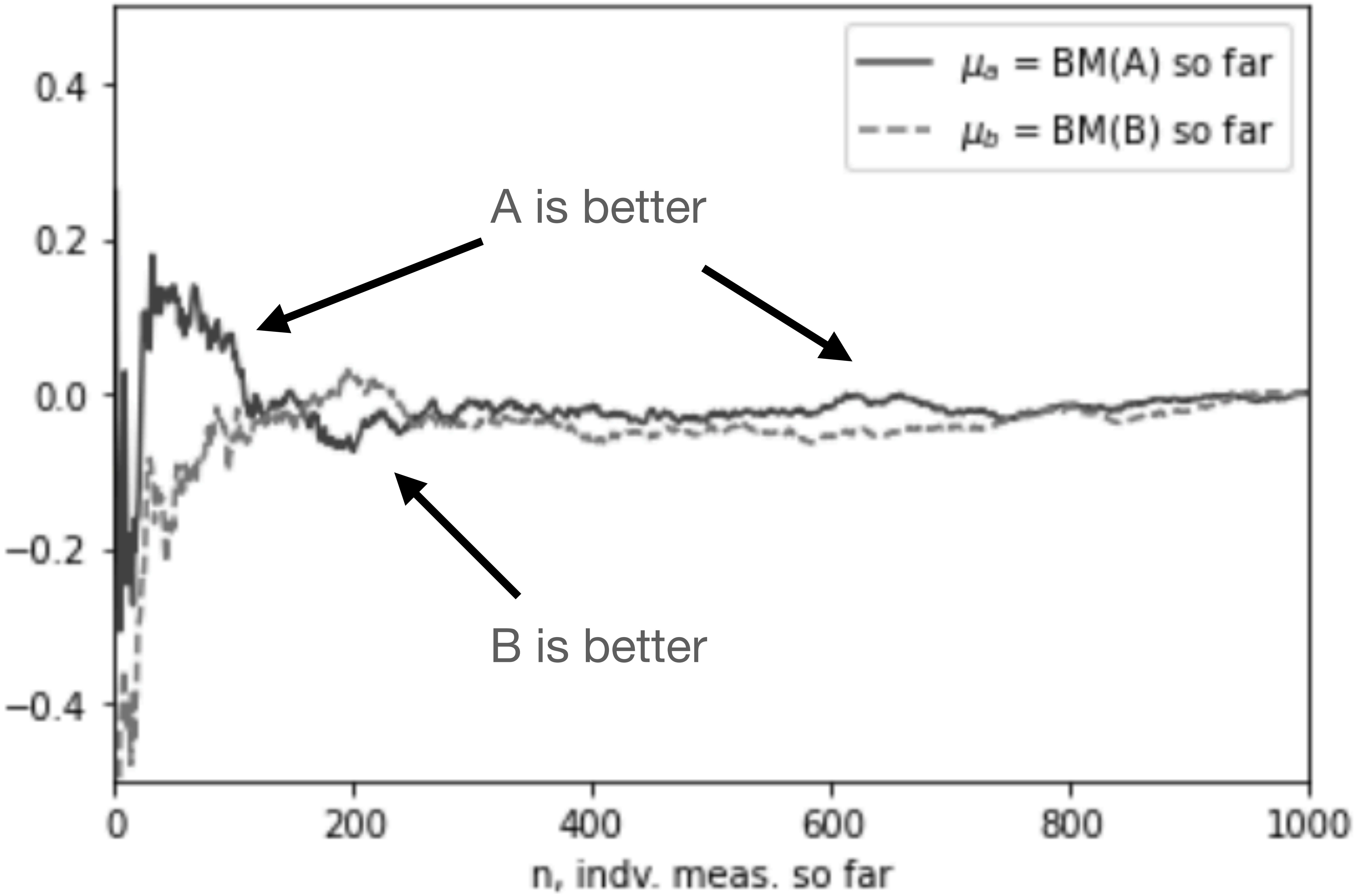
- Note 1: FP/FN errors are more common when BM(B) is closer in value to BM(A).
- Note 2: We're interested in optimizing **business metric, not FP/FN error rates**.
    - We want more revenue, more clicks, less fraud, etc.
- FPR/FNR tell the quality of the experiment. BM tells the quality of the business.

Multi-armed bandits

Optimize the business metric

This is early stopping

- Proposal I: At any point during the experiment, just run whichever version, A or B, has the higher BM.
- Problem: Variation means you could be wrong about which is better and you never get a chance to change your mind.

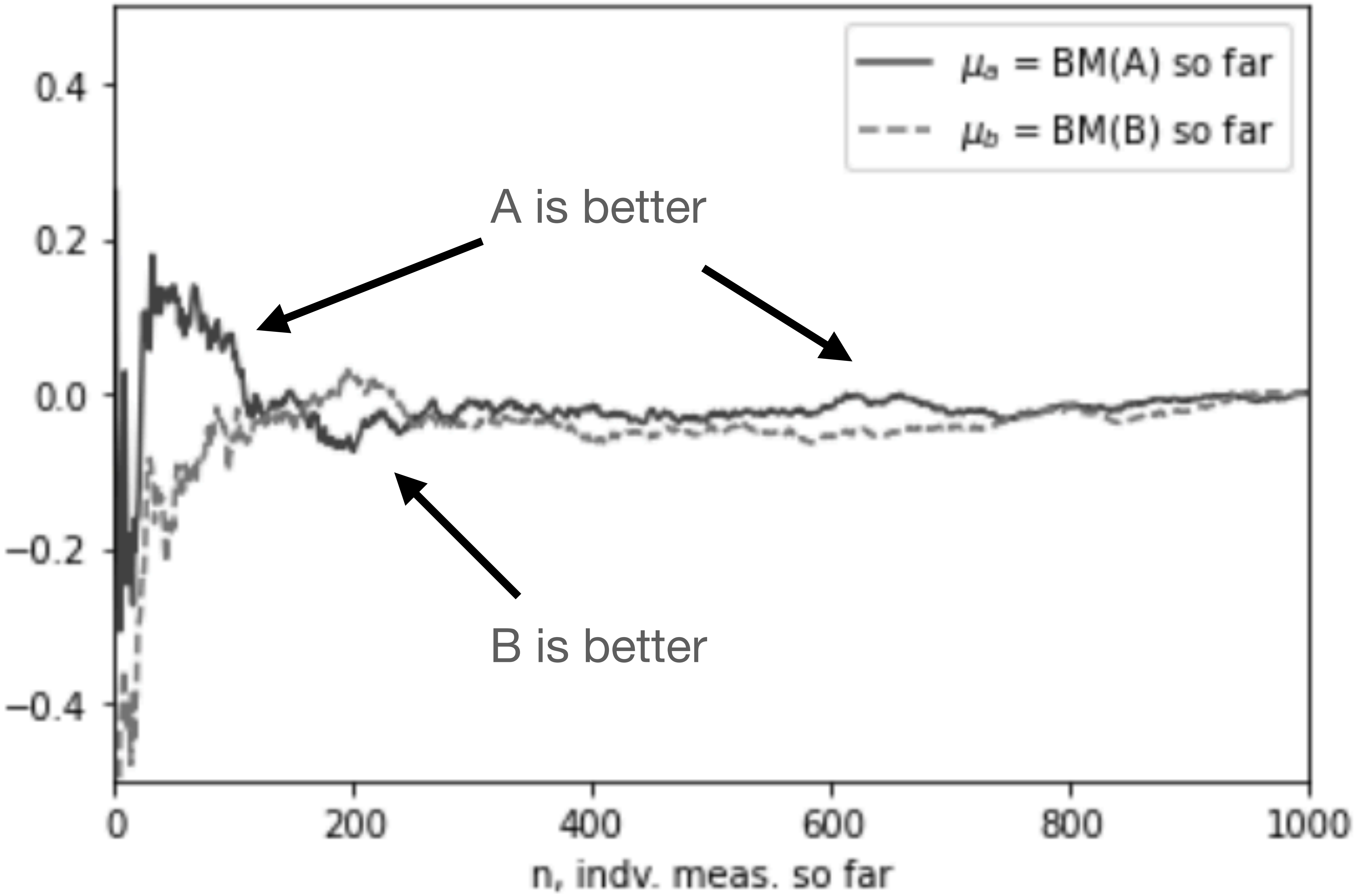


Multi-armed bandits

Optimize the business metric

Not stopping, so not generating FP's

- Proposal II: **Usually** run whichever version, A or B, has the higher BM.
- "usually": 90% of the ind. meas. run the better of A & B
- 10% of time, choose A,B randomly



Multi-armed bandits

Optimize the business metric

- “10% of time, choose A,B randomly”: keeps collecting measurements of “worse” version
  - Allows BM estimate of worse version to continue to vary (maybe later on this will be the better version)
  - Reduces SE of worse version
  - Lower S.E. means more precise comparison of BM’s

Multi-armed bandits

Optimize the business metric

- How does this optimize the business metric?
- At any point during the experiment
  - The one with the better BM-so-far is *probably* the better one
  - You’re probably (90% chance) running the one with the better BM
  - Thus, you’re realizing a better average BM while experimenting

Multi-armed bandits

Epsilon-greedy

- $\varepsilon = 0.10$ ("10% of the time")
- For every individual measurement opportunity:
  - $p_{explore} = \varepsilon$: choose a version, A or B, at random
  - $p_{exploit} = 1 - p_{explore} = 1 - \varepsilon$: run the higher-BM-so-far of A or B
- 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"

You're exploiting the ind. meas. you've collected so far

Multi-armed bandits

Epsilon-greedy: An individual measurement

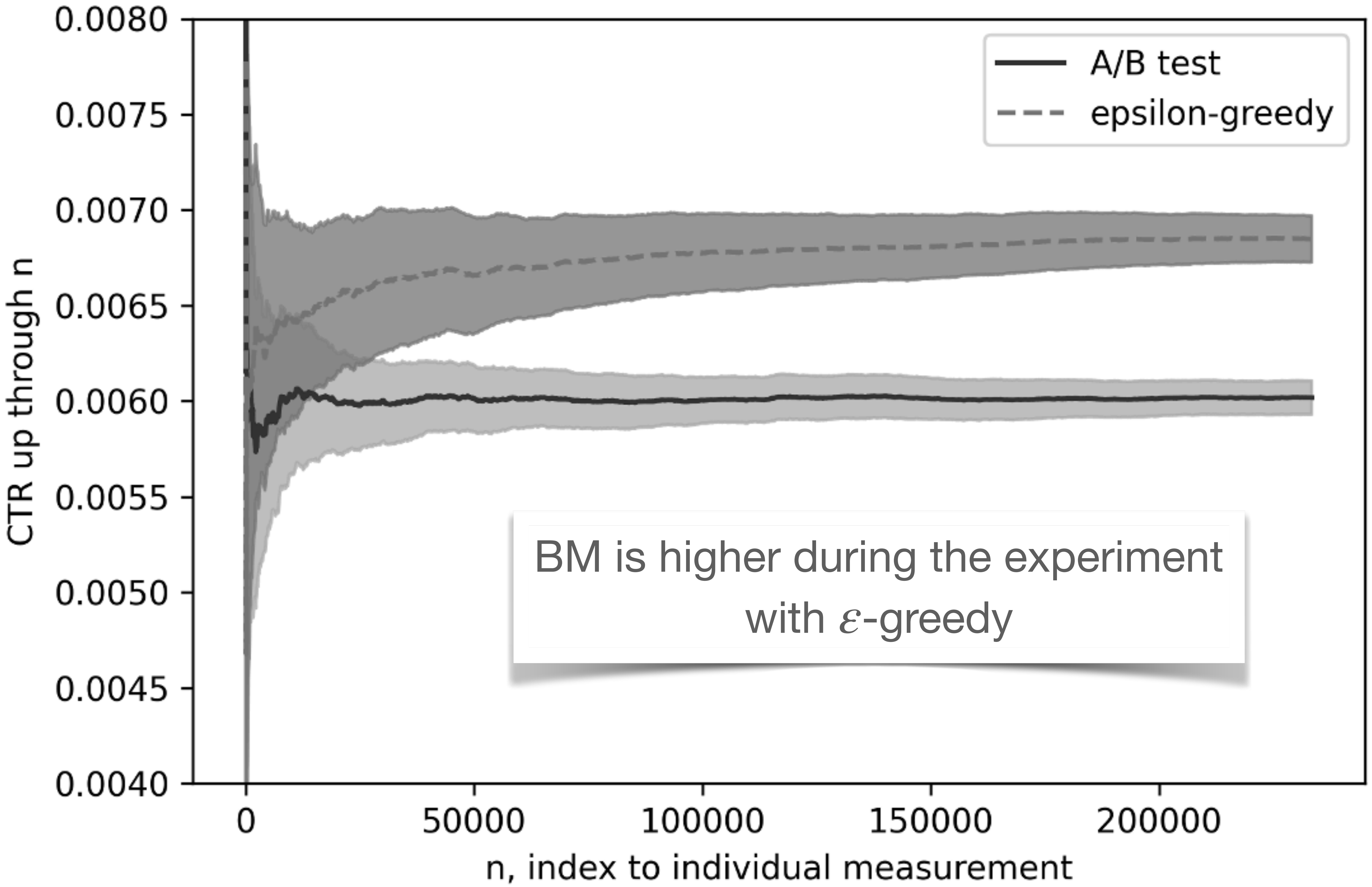
What is the probability of running the better version?

P{FP so far} = Probability that the version with the better BM-so-far is actually the worse verison

	Better version	Worse version
Exploit	0.90 X (1 - P{FP so far})	0.90 X P{FP so far}
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 FPR/FNR” 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
- You could use the 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 individual measurements
- 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} individual measurement: $\varepsilon_n \propto 1/n$
- 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?

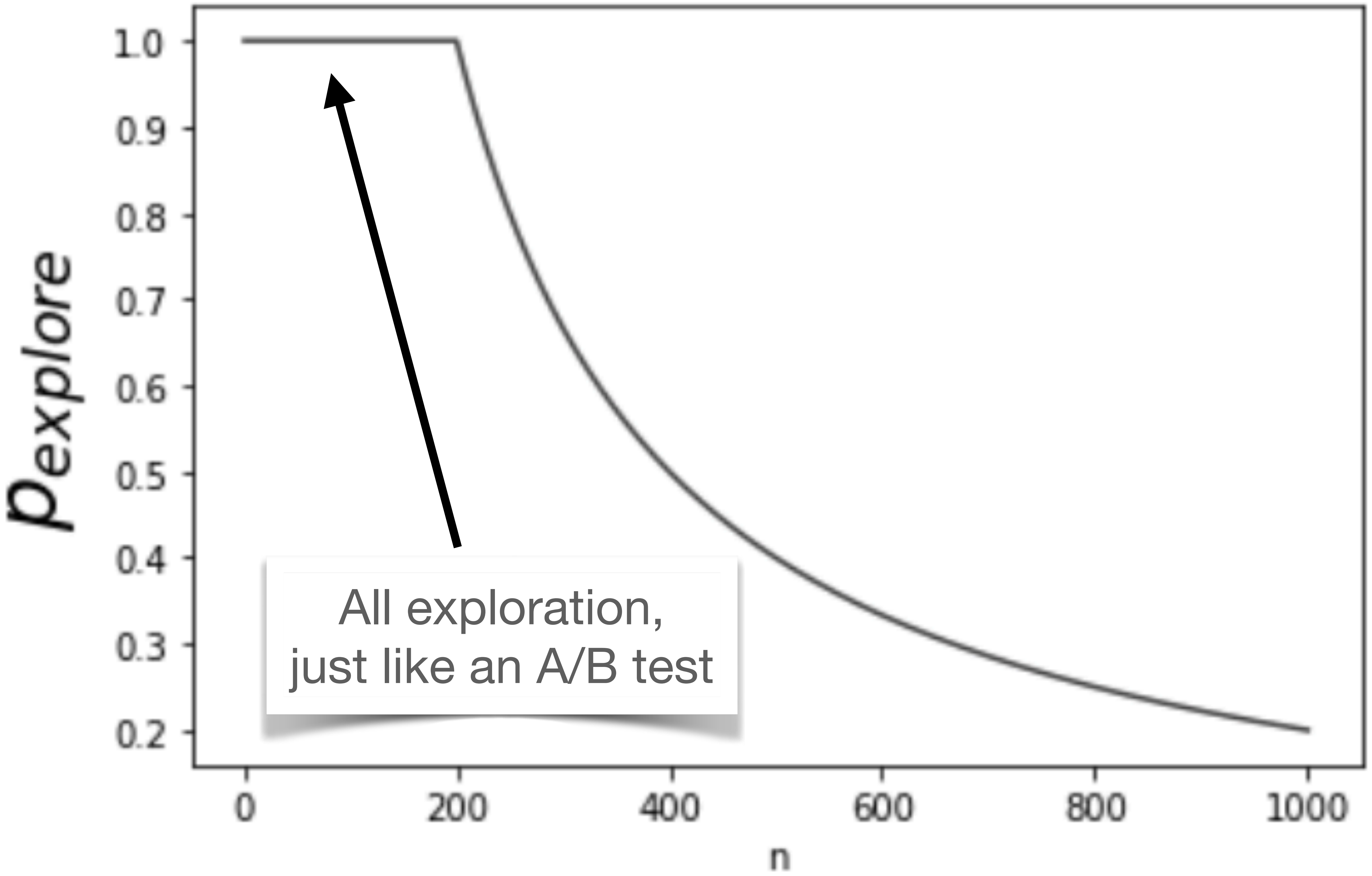
Will a larger PS make this experiment run for more or less time?

- More precisely:
  - $\varepsilon_n = \dfrac{2c(BM_0/PS)^2}{n}$
- 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}

Multi-armed bandits

Epsilon-greedy: When do you stop?

- Since probability can't be larger than one, practically speaking:
  - $p_{explore} = min(1, \varepsilon_n)$
  - $p_{exploit} = 1 - p_{explore}$



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:
  - $\varepsilon_n = \dfrac{\mathbf{2}c(BM_0/PS)^2}{n}$
- to this:
  - $\varepsilon_n = \dfrac{\mathbf{k}c(BM_0/PS)^2}{n}$
- where k is the number of arms.

k=2, here, just A and B

Sometimes called “k-armed bandit”

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 your BM estimates get better (i.e., SE’s get smaller)
  - Stop: When exploration rate is tiny (not really experimenting any more)