Class 6

The power of two choices

Warm-Up

- A hypothetical professor wants to assign grading to TAs, but she has no memory.
  - She can't remember which problem has already been assigned to which TA
- How can she assign n problems to n TAs, so that no TA ends up grading more than $O\left(\frac{\log n}{\log \log n}\right)$ problems?
- Think about how you might do better…
  - It's okay to ask each TA what their current workload is, but you can't remember that info for very long.

Announcements

- HW3 due Friday!

Recap I

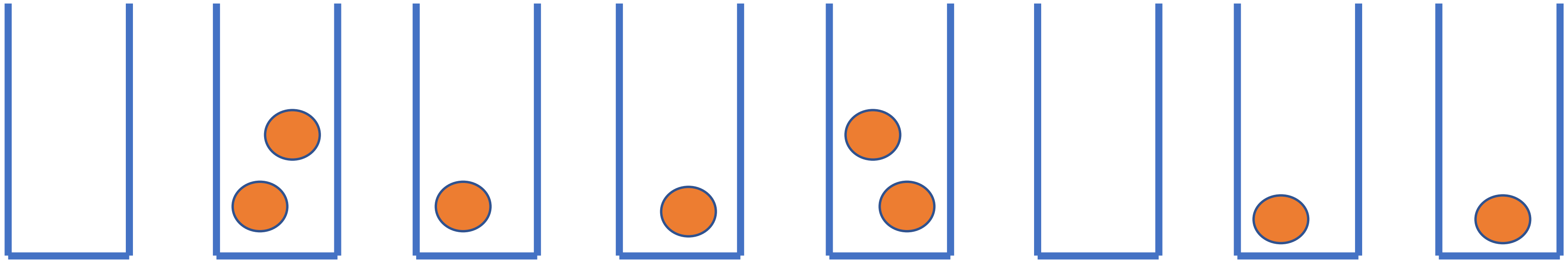
- Balls and bins!
- Powerful tool: Poissonization (Poissonification?)
- $X \sim Poi(\lambda)$:
  - $\Pr[X = k] = \frac{e^{-\lambda}\lambda^k}{k!}$
  - $\mathrm{E}[X] = \mathrm{Var}[X] = \lambda$
  - $\Pr[|X - \lambda| \geq c] \leq 2\exp\left(\frac{-c^2}{2(c+\lambda)}\right)$

Recap II

- If you drop $k \sim Poi(n)$ balls into m bins, then:
  - Let X_i = #(Balls in bin i)
  - $X_i \sim Poi\left(\frac{n}{m}\right)$
  - The X_i are all independent
- "Poissonization":
  - #(Balls in bin i when you drop n balls into m bins) $\approx X_i$
  - Work with the X_i instead.

Recap III: Maximum Load

- n balls into n bins.
- Max load is $\Theta\left(\frac{\log n}{\log\log n}\right)$



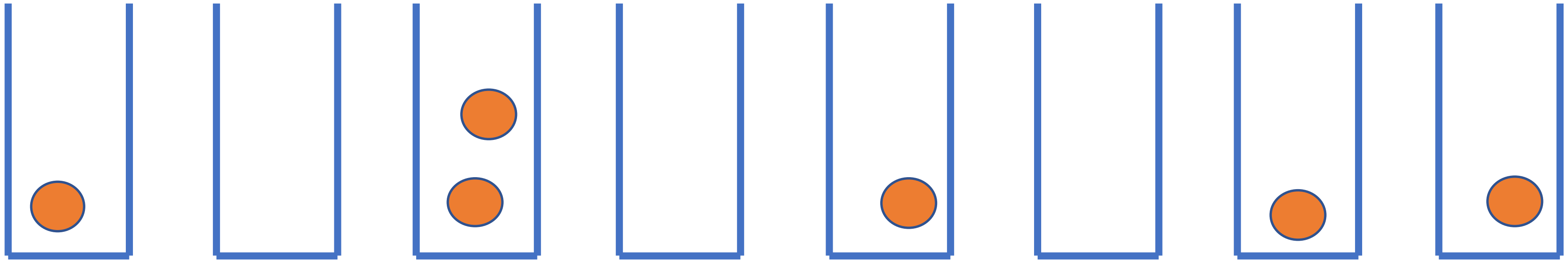
Questions?

Quiz, Mini-lectures, Warm-Up?

- **Solution to warm-up:** assign each problem to a random CA independently!

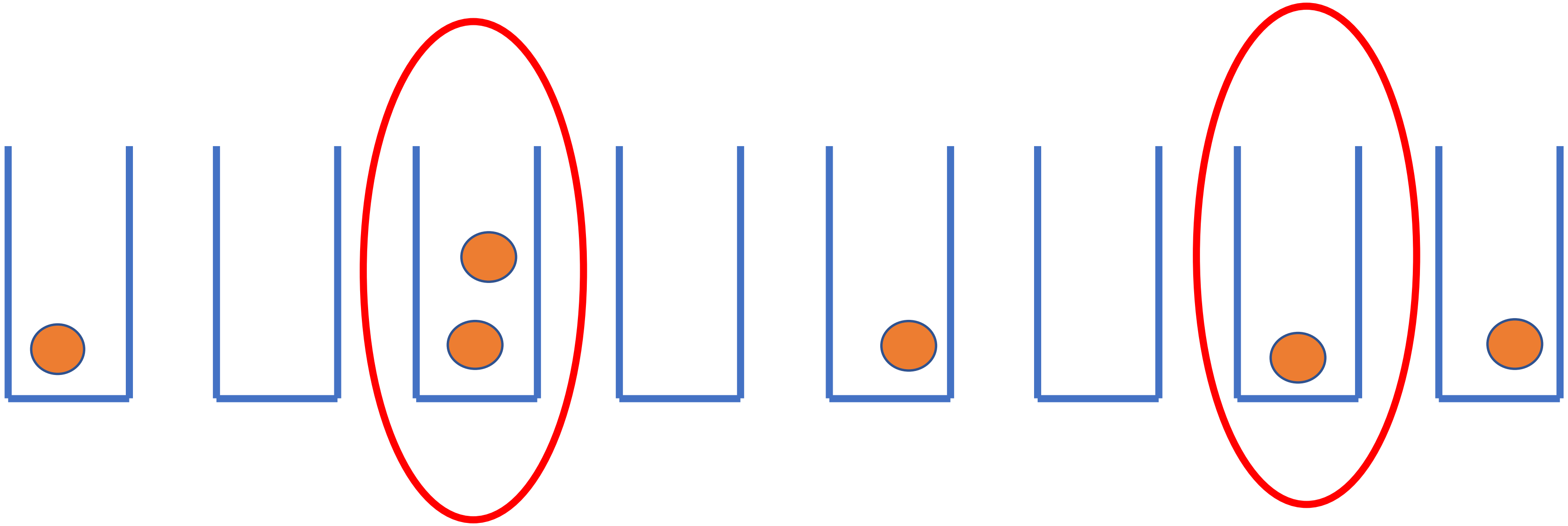
Today: The power of two choices

- Drop n balls into n bins.
- For each ball, pick two bins at random.
- The ball goes in the less-full bin. (Break ties arbitrarily).



Today: The power of two choices

- Drop n balls into n bins.
- For each ball, pick two bins at random.
- The ball goes in the less-full bin. (Break ties arbitrarily).



Today: The power of two choices

- Drop n balls into n bins.
- For each ball, pick two bins at random.
- The ball goes in the less-full bin. (Break ties arbitrarily).

The power of two choices

- n balls into n bins, completely randomly:
  - Max load is $\Theta\left(\frac{\log n}{\log \log n}\right)$

- n balls into n bins, according to the "pick two" scheme:
  - Max load is $\Theta(\log \log n)$

Exponentially smaller!

This is useful, for example, when trying to efficiently assign jobs to processors and wanting to balance the loads.

(Or assigning grading when you have no memory)

For the rest of today…

- We'll analyze this!

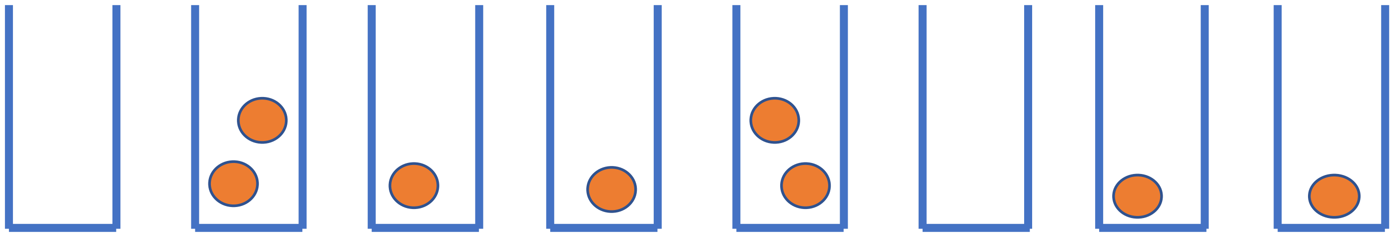
Group Work

- Notation:

$$\beta_2 = \frac{n}{2} \qquad \beta_i = \frac{\beta_{i-1}^2}{n} \text{ for all } i > 2$$

$$\frac{n}{2} \qquad \frac{\left(\frac{n}{2}\right)^2}{n} = \frac{n}{4} \qquad \frac{\left(\frac{n}{4}\right)^2}{n} = \frac{n}{16} \qquad \ldots$$

$B(i,t)$ = number of bins with $\geq i$ balls after step t



This is step $t = 8$

$B(1,8) = 6$

$B(2,8) = 2$

$B(3,8) = 0$

Group Work

Intuition for why this works

Also on handout/website!

Definitions:

$$\beta_2 = \frac{n}{2} \qquad \beta_i = \frac{\beta_{i-1}^2}{n}$$

$B(i,t)$ = number of bins with $\geq i$ balls after step t

1. Explain why $B(2,t) \leq \beta_2$ for all t.

2. Show that

$$\Pr\{\text{Ball } i \text{ is the} \geq \text{3rd ball to land in its bin}\} \leq \left(\frac{B(2,i-1)}{n}\right)^2 \leq \frac{\beta_2^2}{n^2},$$

for all i.

3. Show that, for all t,

$$\mathbb{E}[B(3,t)] \leq \beta_3.$$

← See hint on handout!

4. **Suppose** that $B(3,t) \leq \beta_3$ for all t. That is, suppose that the thing that you showed in expectation before actually held. Show that, for all t,

$$\mathbb{E}[B(4,t)] \leq \beta_4.$$

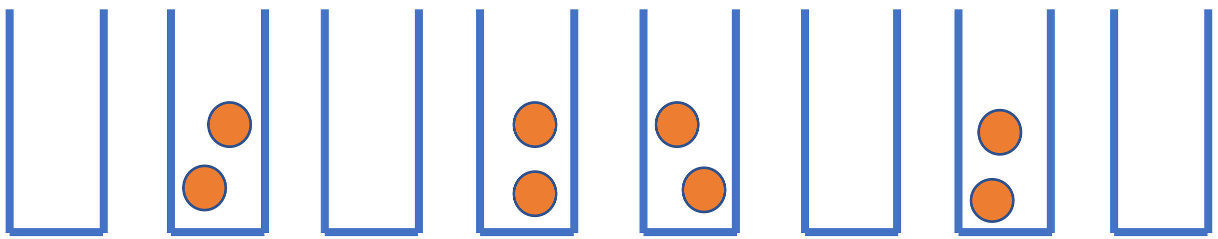
5. Suppose that, for some i^*, $\mathbb{E}[B(i^*,t)]$ was very small (say, less than 0.00001) for all t. Explain why the max load is $\leq i^*$ with high probability.
   Hint: *Markov's inequality*

6. **Suppose** that the logic above continued, and you could show that $\mathbb{E}[B(i,t)] \leq \beta_i$ for all t. Show that, which high probability, the max load is at most $O(\log\log n)$.
   Hint: *Come up with a closed form for β_i. At what point does β_i become way less than 1? Then use the previous part.*

1. Explain why $B(2,t) \leq \beta_2$ for all t.

Solutions: Question 1

- $\beta_2 = n/2$.
- There can't be more than 2 buckets with $\geq n/2$ balls in them (since there's only n balls total).
- So $B(2,t) \leq B(2,n) \leq \beta_2$



This is step $t = 8$

Definitions:

$$\beta_2 = \frac{n}{2} \qquad \beta_i = \frac{\beta_{i-1}^2}{n}$$

$B(i,t)$ = number of bins with $\geq i$ balls after step t

Solutions: Question 2

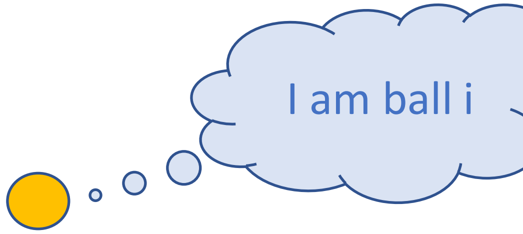
2. Show that

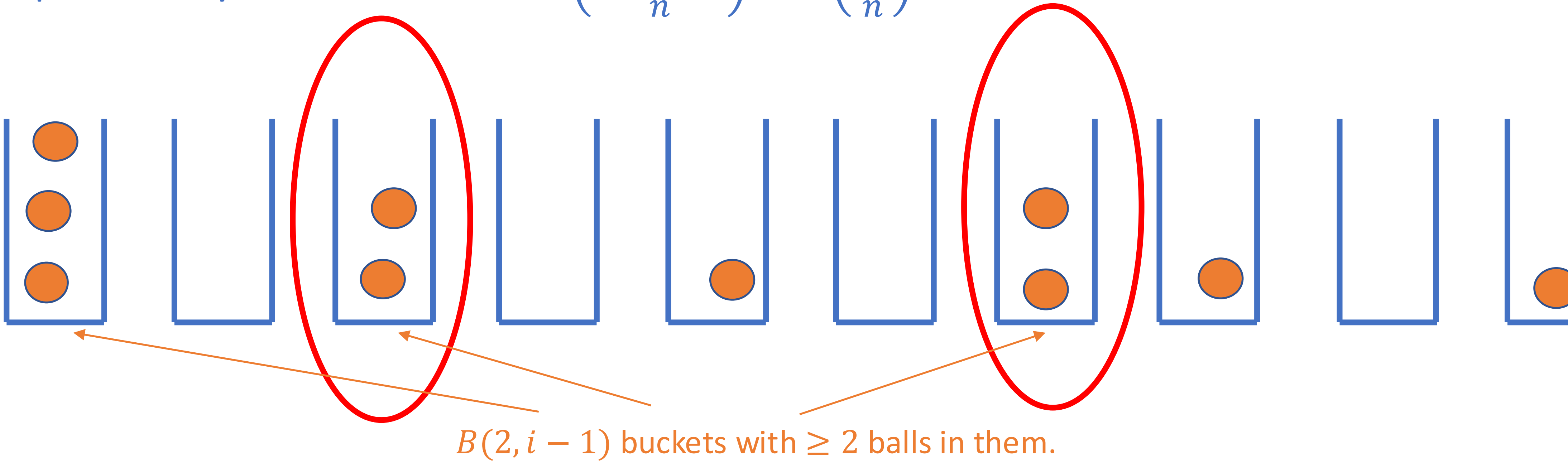
$$\Pr\{\text{Ball } i \text{ is the} \geq \text{3rd ball to land in its bin}\} \leq \left(\frac{B(2,i-1)}{n}\right)^2 \leq \frac{\beta_2^2}{n^2},$$

for all i.

Part 1: $B(2,i) \leq \beta_2$ for all i

- Probability that ball i is the $\geq$ 3rd ball in its bucket:
  - For that to happen, we must choose two buckets with at least 2 balls in them.
  - The probability of that is at most $\left(\frac{B(2,i-1)}{n}\right)^2 \leq \left(\frac{\beta_2}{n}\right)^2$





3. Show that, for all t,

$$\mathbb{E}[B(3,t)] \leq \beta_3.$$

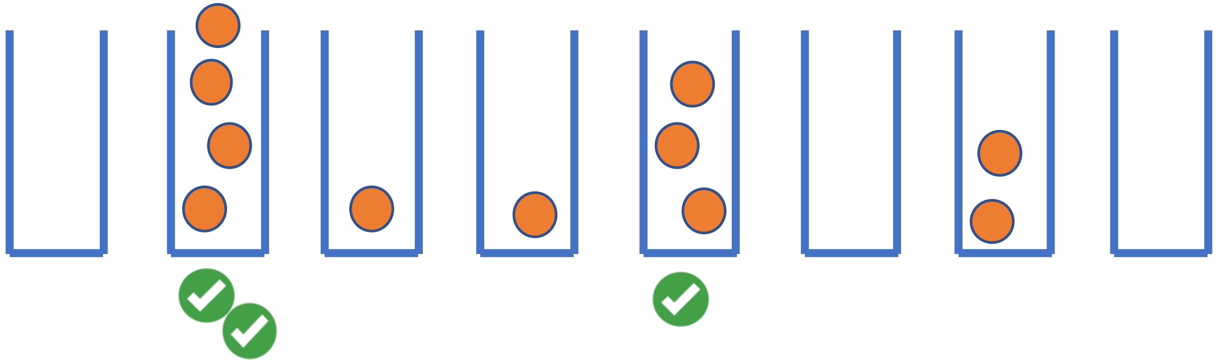
Solutions: Question 3

$$\mathbb{E}[B(3,t)] \leq \mathbb{E}\left[\sum_{t=1}^{t} 1\{\text{ball t is} \geq 3^{rd} \text{ in bucket}\}\right]$$

$$= \sum_{t=1}^{t} Pr\{\text{ball t is} \geq 3^{rd} \text{ in bucket}\}$$

$$\leq \sum_{t=1}^{n} \left(\frac{\beta_2}{n}\right)^2 = \frac{\beta_2^2}{n} = \beta_3$$

This is because the number of bins with at least 3 balls is at most the number of balls that were at least 3rd in their bin.



Definitions:

$$\beta_2 = \frac{n}{2} \qquad \beta_i = \frac{\beta_{i-1}^2}{n}$$

$B(i,t)$ = number of bins with $\geq i$ balls after step t

4. **Suppose** that $B(3,t) \le \beta_3$ for all t. That is, suppose that the thing that you showed in expectation before actually held. Show that, for all t,

$$\mathbb{E}[B(4,t)] \le \beta_4.$$

Solutions: Question 4

$$\mathbb{E}[B(4,t)] \le \mathbb{E}\left[\sum_{t=1}^{t} 1\{\text{ball t is} \ge 4th \text{ in bucket}\}\right]$$

$$= \sum_{t=1}^{t} Pr\{\text{ball t is} \ge 4th \text{ in bucket}\}$$

$$\le \sum_{t=1}^{n} \left(\frac{\beta_3}{n}\right)^2 = \frac{\beta_3^2}{n} = \beta_4$$

This is because the number of bins with at least 4 balls is at most the number of balls that were at least 4th in their bin.

*Note: As per the instructions in the question, we are ignoring anything about conditioning on the event that $B(3,t) \le \beta_3$

5. Suppose that, for some i^*, $\mathbb{E}[B(i^*, t)]$ was very small (say, less than 0.00001) for all t. Explain why the max load is $\leq i^*$ with high probability.

Hint: *Markov's inequality*

Solutions: Question 5

- Say $\mathbb{E}B(i^*, n) \leq 0.001$ for all i.
- By Markov, $\Pr[\ B(i^*, n) \geq 1] \leq 0.001$
- $B(i^*, n)$ is an integer $\Rightarrow \Pr[B(i^*, n) = 0] \geq 0.999$
- Whp, there are zero bins with more than i^* balls
  - aka max load $\leq i^*$

Solutions: Question 6

6. **Suppose** that the logic above continued, and you could show that $\mathbb{E}[B(i,t)] \leq \beta_i$ for all t. Show that, which high probability, the max load is at most $O(\log\log n)$.
Hint: *Come up with a closed form for* β_i. *At what point does* β_i *become way less than 1? Then use the previous part.*

- $\beta_i = \dfrac{n}{2^{2^{i-2}}}$

- You can see this by doing out a bunch and guessing the pattern:
  - $\beta_2 = \frac{n}{2}$
  - $\beta_3 = \frac{1}{n}\left(\frac{n}{2}\right)^2 = \frac{n}{2^2}$
  - $\beta_4 = \frac{1}{n}\left(\frac{n}{2^2}\right)^2 = \frac{n}{2^{2^2}}$
  - $\beta_5 = \frac{1}{n}\left(\frac{n}{2^{2^2}}\right)^2 = \frac{n}{2^{2^3}}$
  - $\beta_6 = \frac{1}{n}\left(\frac{n}{2^{2^3}}\right)^2 = \frac{n}{2^{2^4}}$

(And formally you can prove it by induction.)

Solutions: Question 6

6. **Suppose** that the logic above continued, and you could show that $\mathbb{E}[B(i,t)] \leq \beta_i$ for all t. Show that, which high probability, the max load is at most $O(\log\log n)$.
Hint: Come up with a closed form for β_i. At what point does β_i become way less than 1? Then use the previous part.

- $\beta_i = \dfrac{n}{2^{2^{i-2}}}$

- By the previous part, if $\mathbb{E}B(i^*, n) \leq 0.001$, then max load $\leq i^*$ whp.
- Solve $\beta_i \leq 0.001$:

$$\frac{n}{2^{2^{i-2}}} \leq 0.001$$

$$n \leq 0.001 \cdot 2^{2^{i-2}}$$

$$\log\log n \leq \log\left(2^{i-2} + \log(0.001)\right) \approx i - 2$$

- So if $i^* \approx \log\log n + 2$, we should be good!

If you take $i^* = \log\log n + 3$, you can check that $\beta_{i^*} = \frac{1}{n} \ll 0.001$.

What we've just seen

- Some intuition for why:

With power-of-two-choices, max load $\leq O(\log \log n)$

- **But!** We were cheating!
- We can't pretend that $\mathbb{E}B(i,t) \leq \beta_i \Rightarrow B(i,t) \leq \beta_i$

How to fix this?

Here’s the outline for a fixed argument

Also sketched on handout!

WARNING: This is incorrect in a few ways.

This “2” is new

1. Define $\beta_4 = \frac{n}{4}, \ \beta_i = \frac{\mathbf{2}(\beta_{i-1})^2}{n}$

2. Argue by induction that, with probability $\geq 1 - \frac{i}{n^2}$, $B(i,n) \leq \beta_i$:

Number of bins with $\geq i$ balls after all n are tossed

- Base case for $i = 4$ holds by definition (with probability 1).
- Assuming that $B(i-1,n) \leq B_{i-1}$ (which holds with prob $\geq 1 - \frac{i-1}{n^2}$ by induction), the same logic as before implies

$$\mathbb{E}\left[\sum_{t=1}^{n} 1[\text{ball t is} \geq i'\text{th in its bucket}]\right] \leq \frac{\beta_{i-1}^2}{n}$$

- A Chernoff bound implies

This is β_i, and also twice the expectation

$$\Pr\left[\sum_{t=1}^{n} 1[\text{ball t is} \geq i'\text{th in its bucket}] > 2\frac{\beta_{i-1}^2}{n}\right] \leq \exp\left(-\frac{\beta_i}{3}\right)$$

As before, $B(i,n) \leq$ this

2. Argue by induction that, with probability $\geq 1 - \frac{i}{n^2}$, $B(i,n) \leq \beta_i$:

Number of bins with $\geq i$ balls after all n are tossed

- Base case for $i = 4$ holds by definition (with probability 1).
- Assuming that $B(i-1,n) \leq B_{i-1}$ (which holds with prob $\geq 1 - \frac{i-1}{n^2}$ by induction), the same logic as before implies

$$\mathbb{E}\left[\sum_{t=1}^{n} 1[\text{ball t is} \geq i'\text{th in its bucket}]\right] \leq \frac{\beta_{i-1}^2}{n}$$

- A Chernoff bound implies

This is β_i, and also twice the expectation

$$\Pr\left[\quad\ldots\quad B(i,n) \quad\ldots\quad > 2\frac{\beta_{i-1}^2}{n}\right] \leq \exp\left(-\frac{\beta_i}{3}\right)$$

2. Argue by induction that, with probability $\geq 1 - \frac{i}{n^2}$, $B(i,n) \leq \beta_i$:

Number of bins with $\geq i$ balls after all n are tossed

Prob of this is $\leq \frac{i-1}{n^2}$, by induction

- Assuming that $B(i-1,n) \leq B_{i-1}$, a Chernoff bound implies

$$\Pr[B(i,n) > \beta_i] \leq \exp\left(-\frac{\beta_i}{3}\right)$$

- If $\beta_i \geq 6 \log n$, $\exp\left(-\frac{\beta_i}{3}\right) \leq 1/n^2$.
- Union bound with the event that $B(i-1,n) > \beta_{i-1}$, and we see $\Pr[B(i,n) > \beta_i] \geq \frac{i}{n^2}$

So as long as $\beta_i \geq 6 \log n$, we can establish the inductive hypothesis for the next round!

Now we know that $B(i,n) \leq \beta_i$ with probability at least $1 - i/n^2$, as long as $\beta_i \geq 6\log n$

3. Choose i^* so that $\beta_{i^*} \geq 6\log n$

Then the argument above shows that, whp, $B(i,n) \leq \beta_i \ \forall i \leq i^*$

You can check that $\beta_{i^*} \approx 6\log n$ when $i^* = \Theta(\log\log n)$

This part is not the problem!

3. We conclude that, whp, $B(i,n) \leq \beta_i$ for all $i \leq i^* = \Theta(\log\log n)$

As before, this implies that the max load is $\Theta(\log\log n)$

Group Work

- What was wrong with this argument/sketch?
- There are at least two or three major problems
    - depending on what you count as “major”

Three problems

1. Can't apply the Chernoff bound – the random variables are not independent!
2. The end of the argument doesn't make any sense! We showed that $B(i, n) \leq \beta_i$ whenever $i \leq i^*$, but $\beta_{i^*} \approx 6 \log n$, not 0.001.
   - So there are still about $6 \log n$ buckets with at least i^* balls in them.
   - (It is true that $i^* = \Theta(\log \log n)$ though).
3. We are not being careful about the conditioning.

It turns out that all these problems can be fixed!

We won't go over these fixes in class and you aren't responsible for them. Check out the survey linked on the course website if you are curious!

- Short versions:

1. Can't apply the Chernoff bound on $\sum_t X_t$ if the X_t aren't independent!
   - We can argue that $\mathbb{E}[X_t | X_1, \ldots, X_{t-1}] \leq \frac{\beta_{i-1}^2}{n}$ still holds.
   - This turns out to be enough! We'll see why later in the course ☺
2. The end of the argument doesn't make any sense!
   - If you go one more step to $i^* + 1$, it will make sense.
3. We are not being careful about the conditioning.
   - Be more careful about the conditioning.

Next time!

- Metric Embeddings!!