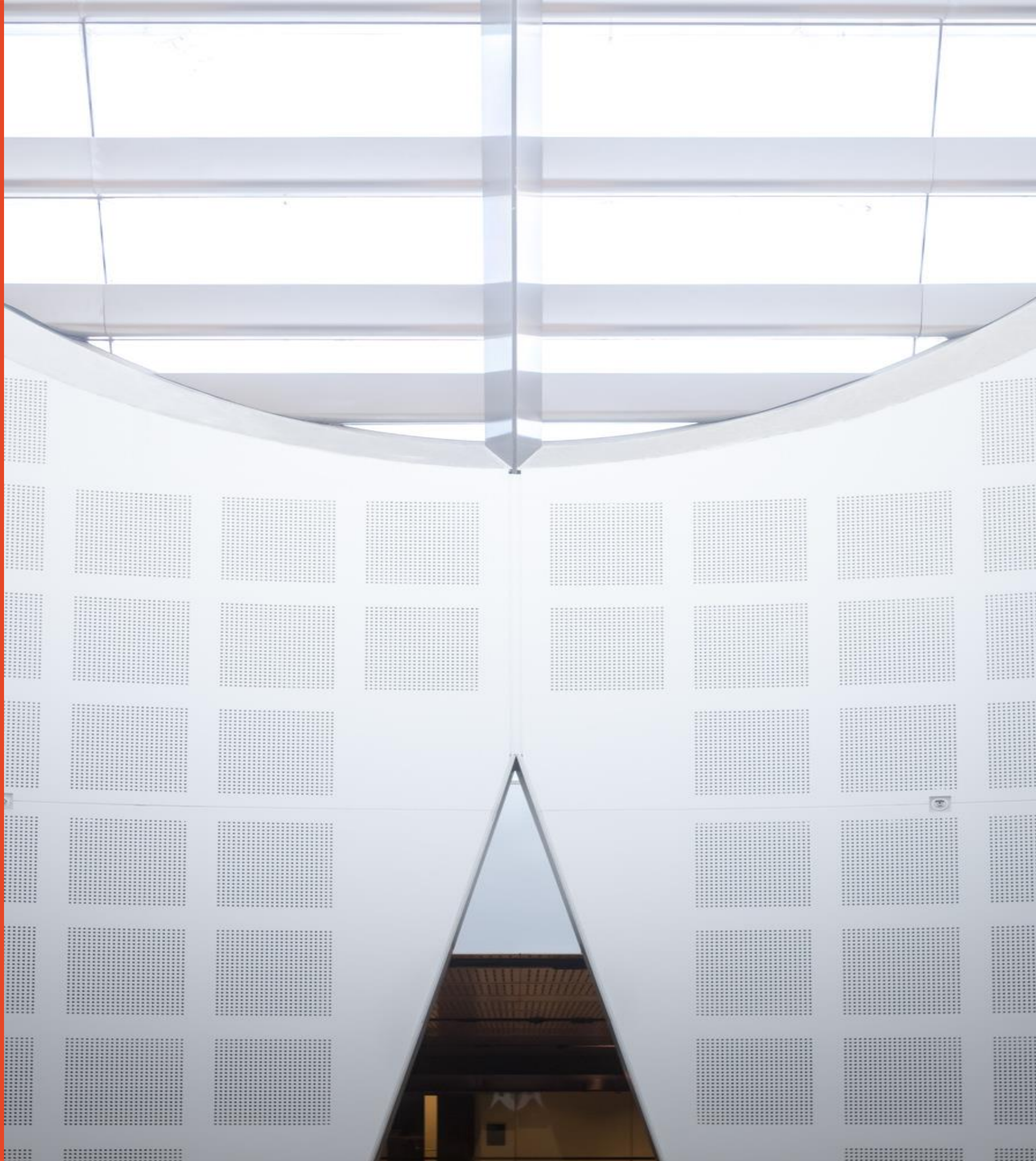
COMPx270: Randomised and Advanced Algorithms
Lecture 3: Balls in Bins

Clément Canonne

School of Computer Science

THE UNIVERSITY OF SYDNEY



A question 🎂

There are quite a few people in the classroom right now. What are the odds two of you (at least) have the same birthday?

A question 🎂

Theorem. (The 🎂 paradox) If you gather 23 people in a room, then with probability at least 50% a pair will sharing their birthday.

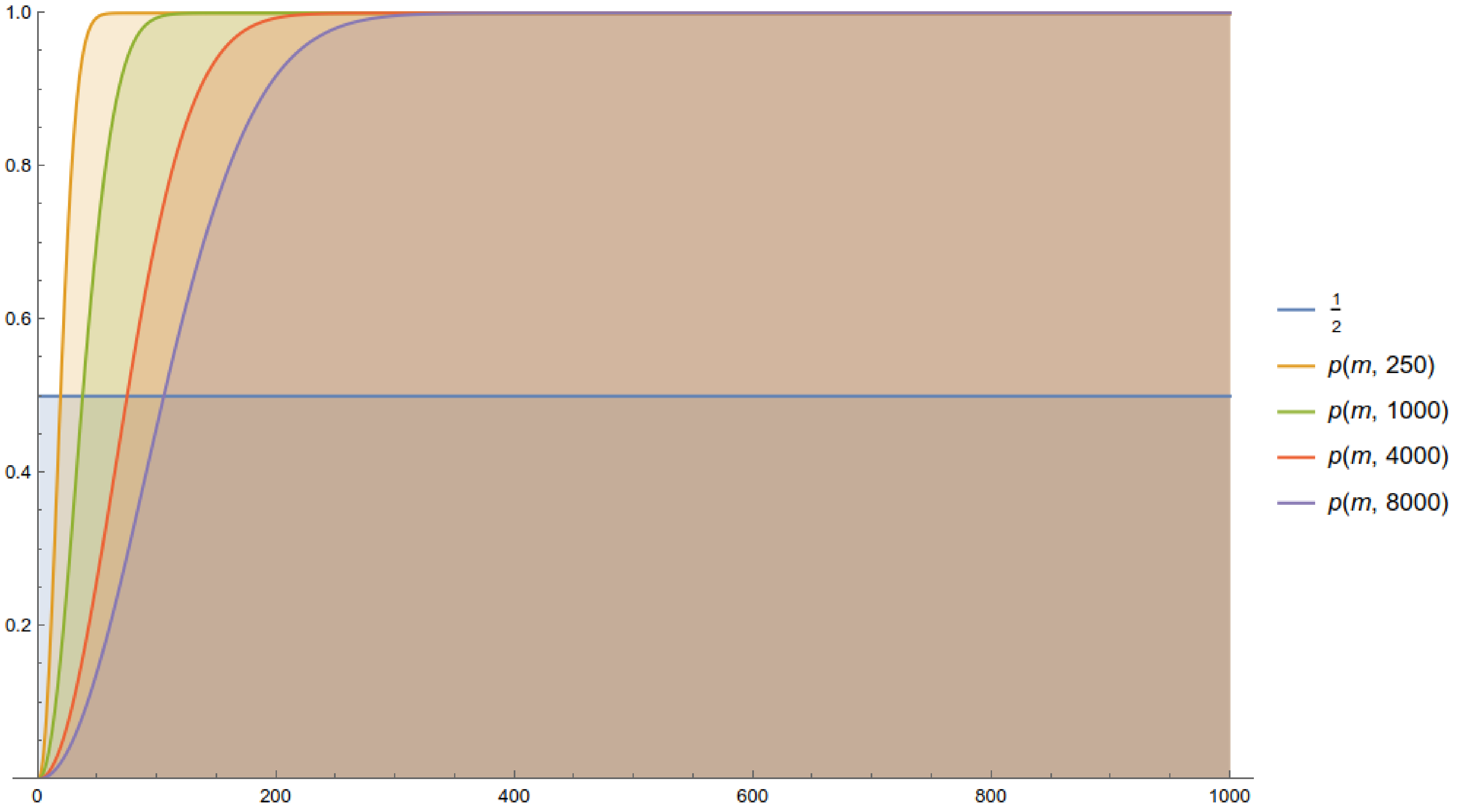
An answer? 🎂

Theorem. If you gather m people and give each a number uniform between 1 and n, then the probability p(m,n) that at least two have the same number is...

$$p_{m,n} = 1 - \frac{n!}{n^m (n-m)!} = 1 - \frac{m!}{n^m}\binom{n}{m}$$

Proof.

```
DiscretePlot[{1/2, p[m, 250], p[m, 1000], p[m, 4000], p[m, 8000]}, {m, 1, 1000}, PlotRange → {0, 1}, PlotLegends → "Expressions"]
```



Let's start simple: m=2 ⚾ ⚾

Now, for large values of "2"...

C: number of collisions when throwing m ⚾ into n 🗑️ . What is

$$c(m,n) = E[C]?$$

Number of collisions

... and what is Var[C]?

Number of collisions

... and what is Var[C]?

$$C = \sum_{\ell=1}^{\binom{m}{2}} Y_\ell$$

Y_ℓ: 0 if no collision on ℓ^{th} pair, 1 o/w

$$\mathbb{E}C = \frac{\binom{m}{2}}{n}$$

If C was Binomial

$$\left(C \overset{?}{\sim} \mathrm{Bin}\left(\binom{m}{2}, \frac{1}{n}\right) \right)$$

$$\mathrm{Var}\, C = \binom{m}{2} \cdot \frac{1}{n}\left(1 - \frac{1}{n}\right)$$

C is not Binomial, b/c $Y_1, \ldots, Y_\ell$ not indep^t.

so...

$$\mathrm{Var}\, C = \mathbb{E}[C^2] - \mathbb{E}[C]^2$$

$\mathbb{E}[C^2]$: ?

$\mathbb{E}[C]$: known

$M = \binom{m}{2}$

$$\mathrm{Var}[C] = \binom{m}{2}\frac{1}{n}\left(1-\frac{1}{n}\right)$$

Number of collisions

$$\mathbb{E}[C^2] = \mathbb{E}\left[\left(\sum_{\ell=1}^{M} Y_\ell\right)^2\right] = \mathbb{E}\left[\sum_{\ell=1}^{M}\sum_{k=1}^{M} Y_\ell Y_k\right]$$

(yay)

$$= \sum_{\ell=1}^{n}\sum_{k=1}^{M} \mathbb{E}[Y_\ell Y_k] = ?$$

$$\binom{m}{2}\frac{1}{n} + \binom{m}{2}\binom{m-2}{2}\frac{1}{n^2}$$

$$+ 6\binom{m}{3}\frac{1}{n^2}$$

$$= \binom{m}{2}\frac{1}{n}\left(1-\frac{1}{n}\right) + \binom{m}{2}^2\frac{1}{n^2}$$

annoying distinct of cases

$$\mathrm{Var}\, C = \mathbb{E}[C^2] - \mathbb{E}[C]^2$$

$$= \binom{m}{2}\frac{1}{n}\left(1-\frac{1}{n}\right)$$

\o/

$$\mathrm{Var}[C] = \binom{m}{2}\frac{1}{n}\left(1 - \frac{1}{n}\right)$$

Number of collisions

Now, we can use Chebyshev:

$$\Pr[\, X = 0 \,] \leq 1/2$$

for $m = \Omega(\sqrt{n})$.

$$\mathrm{Var}[C] = \binom{m}{2}\frac{1}{n}\left(1 - \frac{1}{n}\right)$$

Number of collisions

Now, we can use Chebyshev:

$$\Pr[\ X = 0\ ] \leq 1/2$$

for $m = \Omega(\sqrt{n})$. **Is it tight?**

Number of collisions

$$\mathrm{Var}[C] = \binom{m}{2} \frac{1}{n} \left(1 - \frac{1}{n}\right)$$

Now, we can use Chebyshev:

$$\Pr[\, X = 0 \,] \leq 1/2$$

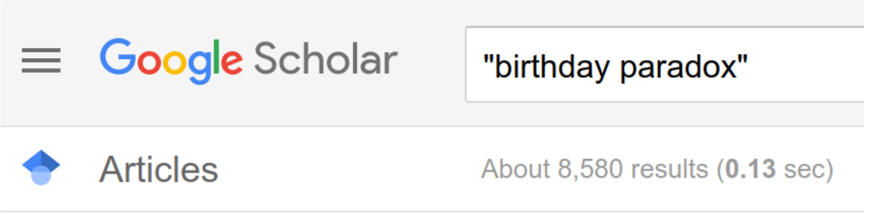
for $m = \Omega(\sqrt{n})$.

By Markov, we also have

$$\Pr[\, X \neq 0 \,] = \Pr[\, X \geq 1 \,] \leq E[X] \leq 1/2$$

for $m = O(\sqrt{n})$.

Applications?



Bounding the variance: is it always that bad?

Two tricks (and even 3).

① $\mathrm{Var} \sum_i X_i \neq \sum_i \mathrm{Var}\, X_i$ in general

But "often" $\mathrm{Var} \sum_i X_i \leq \sum_i \mathrm{Var}\, X_i$

$X_1, \dots, X_n$ "negatively correlated"

② $\mathrm{Var}\, X \leq \mathbb{E}[X^2]$ $\quad$ $\left(\mathrm{Var}\, X = \mathbb{E}[X^2] - \underbrace{\mathbb{E}[X]^2}_{\geq 0} \leq \mathbb{E}[X^2]\right)$

(③ "Pairwise independence" is enough
(will get back to it))

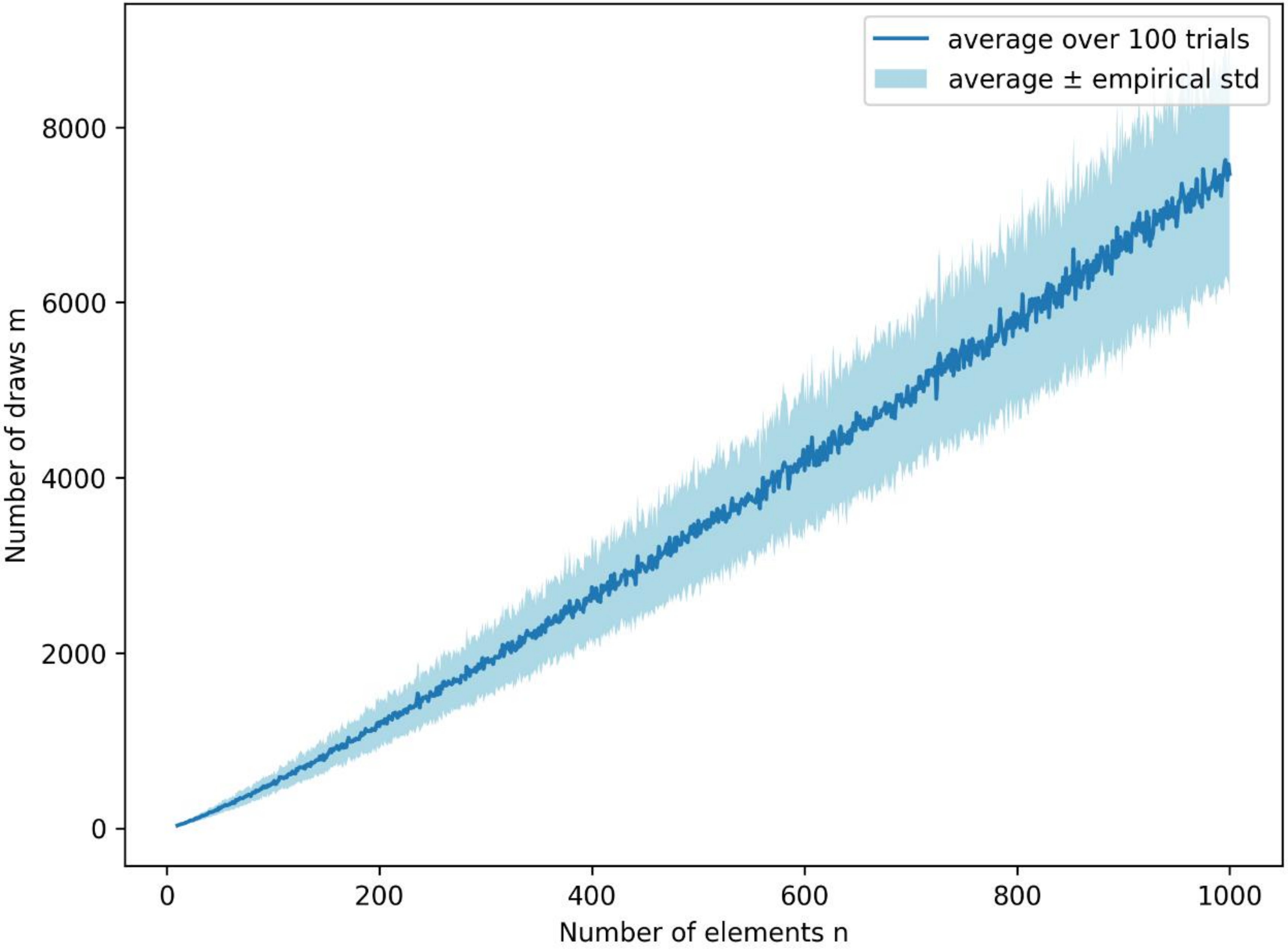
Coverage (Coupon collector)

"What is the expected number of balls $M(n)$ we need to throw before each of the n bins contains at least one ball?"

Coverage (Coupon collector)

"What is the expected number of balls $M(n)$ we need to throw before each of the n bins contains at least one ball?"

- $\Theta(n)$?
- $\Theta(n \log n)$?
- $\Theta(n^2)$?
- Something else?


average over 100 trials
average ± empirical std
Number of draws m
0
2000
4000
6000
8000
0
200
400
600
800
1000
Number of elements n

Coverage (Coupon collector)

"What is the expected number of balls M(n) we need to throw before each of the n bins contains at least one ball?"

Theorem. In expectation, M(n) = Θ(n log n) balls. (Even more precisely: n ln n + O(n).)

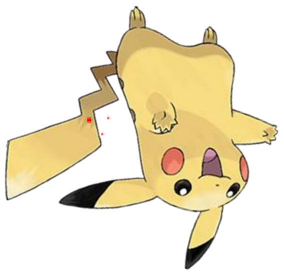
Intuition: throw n balls. $\Pr[\text{no Pikachu}] = \left(1-\frac{1}{n}\right)^n \approx \frac{1}{e}$

(. $\Pr[\exists \text{ Pokemon not seen}] \underset{\text{union}}{\leq} n \cdot \left(1-\frac{1}{n}\right)^n \approx \frac{n}{e}$)

. $\mathbb{E}[\text{Pokemon not seen}] = n\left(1-\frac{1}{n}\right)^n \approx \frac{n}{e}$

n more: $\mathbb{E}[\ ] = \frac{n}{e^2}$. n more: $\frac{n}{e^3}$

Proof.



T_i: conditioned on having $i-1$ Pokemon, what is the # of Pokeballs until I get an i^{th} Pokemon (new)

$T_1 = 1$

$$M(n) = T_1 + \cdots + T_n$$

$$\mathbb{E}[M(n)] = \sum_{i=1}^{n} \mathbb{E}[T_i] = \sum_{i=1}^{n} \frac{n}{n-i+1}$$

$$\Pr[\text{getting } i^{th} \text{ new} \mid \text{I have } i-1 \text{ distinct Pokemon}] = \frac{n-(i-1)}{n}$$

$$= n \sum_{i=1}^{n} \frac{1}{n-i+1}$$

Geometric!

r.v.

$$T_i \sim \mathrm{Geom}\left(\frac{n-i+1}{n}\right) \qquad \mathbb{E}[T_i] = \frac{n}{n-i+1}$$

$$\underset{\text{"}\ell = n-i+1\text{"}}{=} n \cdot \sum_{\ell=1}^{n} \frac{1}{\ell} = nH_n$$

$$= n(\ln n + O(1))$$

What about the variance?

$$\operatorname{Var} M(n) = \operatorname{Var}(T_1 + \dots + T_n)$$

$$\underset{\substack{\text{independent} \\ \text{(magic)}}}{=} \sum_{i=1}^{n} \operatorname{Var} T_i$$

$$= \sum_{i=1}^{n} \frac{n^2}{(n-i+1)^2} \cdot \underbrace{\left(1 - \frac{n-i+1}{n}\right)}_{\le 1}$$

$$\le n^2 \sum_{i=1}^{n} \frac{1}{(n-i+1)^2}$$

$$= n^2 \sum_{\ell=1}^{n} \frac{1}{\ell^2}$$

$$\underset{(n<\infty)}{\le} n^2 \sum_{\ell=1}^{\infty} \frac{1}{\ell^2} = \frac{\pi^2}{6} n^2$$

Geom(p) has variance $\frac{1-p}{p^2}$

Chebyshev

$M(n) = n \log n \pm O(n)$ w.h.p.

Load balancing

"What is the expected number of balls $L(n)$ the fullest of the n bins contains after throwing n balls?"

Load balancing

"What is the expected number of balls L(n) the fullest of the n bins contains after throwing n balls?"

- $\Theta(1)$?
- $\Theta(\log n)$?
- $\Theta(\sqrt{n})$?
- Something else?

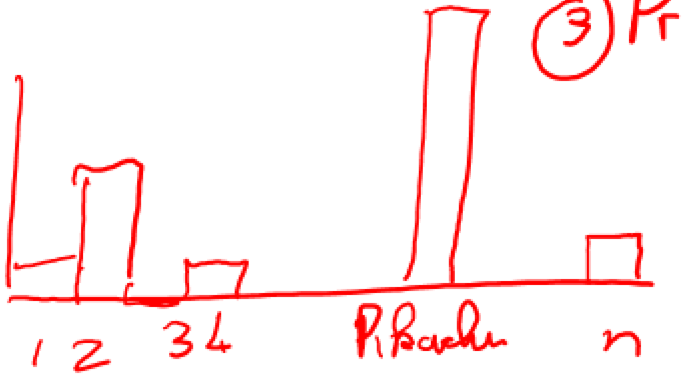
1 $\mathbb{E}[L_i] = 1$

i^{th} Bin

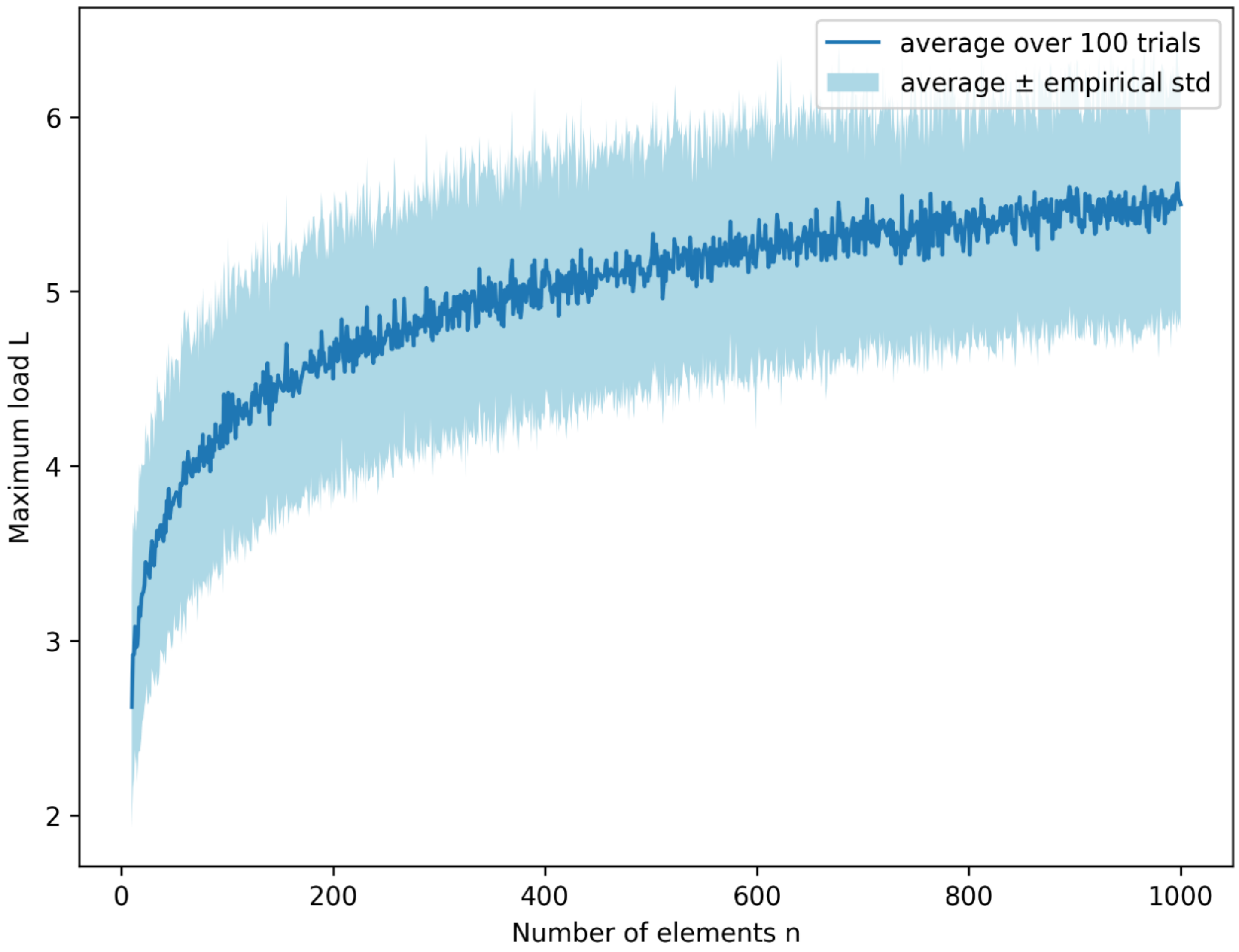
2 $L_i \sim Bin(n, \frac{1}{n})$

$Var[L_i] = n \frac{1}{n}(1 - \frac{1}{n}) \leq 1$

3 $Pr[L_i > 1+\sqrt{n}] \leq \frac{1}{2n}$



$Pr[L(n) \geq 1+\sqrt{n}] \leq 1/2$



average over 100 trials
average ± empirical std
Maximum load L
Number of elements n
2
3
4
5
6
0
200
400
600
800
1000

Load balancing

"What is the expected number of balls L(n) the fullest of the n bins contains after throwing n balls?"

Theorem. The expected maximum load is L(n) = Θ(log n / log log n).

$$\mathbb{E}[L] = \sum_{k=1}^{n} \Pr[L \geq k] \leq n \sum_{k=1}^{n} \Pr[L_1 \geq k]$$

$$L = \max_{1 \leq i \leq n} L_i$$

$$\Pr[L \geq k] = \Pr[\exists i,\ L_i \geq k] \underset{\text{union}}{\leq} \sum_{i=1}^{n} \Pr[L_i \geq k]$$

Proof.

$1 \le k \le n$

$$\Pr[L_1 \ge k] \le \binom{n}{k}\left(\frac{1}{n}\right)^k = \frac{n!}{k!(n-k)!\,n^k}$$

?!

$$\left(\frac{n}{k}\right)^k \le \binom{n}{k} \le \left(\frac{e\cdot n}{k}\right)^k$$

Proof.

$$\Pr[L_1 \geq k] \leq \left(\frac{en}{k}\right)^k \cdot \frac{1}{n^k} = \frac{e^k}{k^k} \quad \checkmark$$

$$\mathbb{E}[L] \leq \sum_{k=1}^{n} \Pr[L \geq k] \leq \sum_{k=1}^{n} \min\left(\frac{ne^k}{k^k}, 1\right)$$

$$\leq \sum_{k=1}^{\ell} 1 + \sum_{k=\ell+1}^{n} \frac{ne^k}{k^k} \leq \ell + \sum_{k=\ell+1}^{\infty} \frac{ne^k}{k^k}$$

$\Pr[L \geq k] \leq \frac{ne^\ell}{k^k}$

$\Pr[L \geq k] \leq 1$

Pick ℓ s.t. $\frac{ne^\ell}{\ell^k} < \frac{1}{2^k}$ for $k \geq \ell+1$

$$\leq \ell + O(1)$$

Need $2^\ell ne^\ell \leq \ell^\ell$

$$\leq O\left(\frac{\log n}{\log\log n}\right) \longleftarrow \ell = \Theta\left(\frac{\log n}{\log\log n}\right) \text{ works}$$

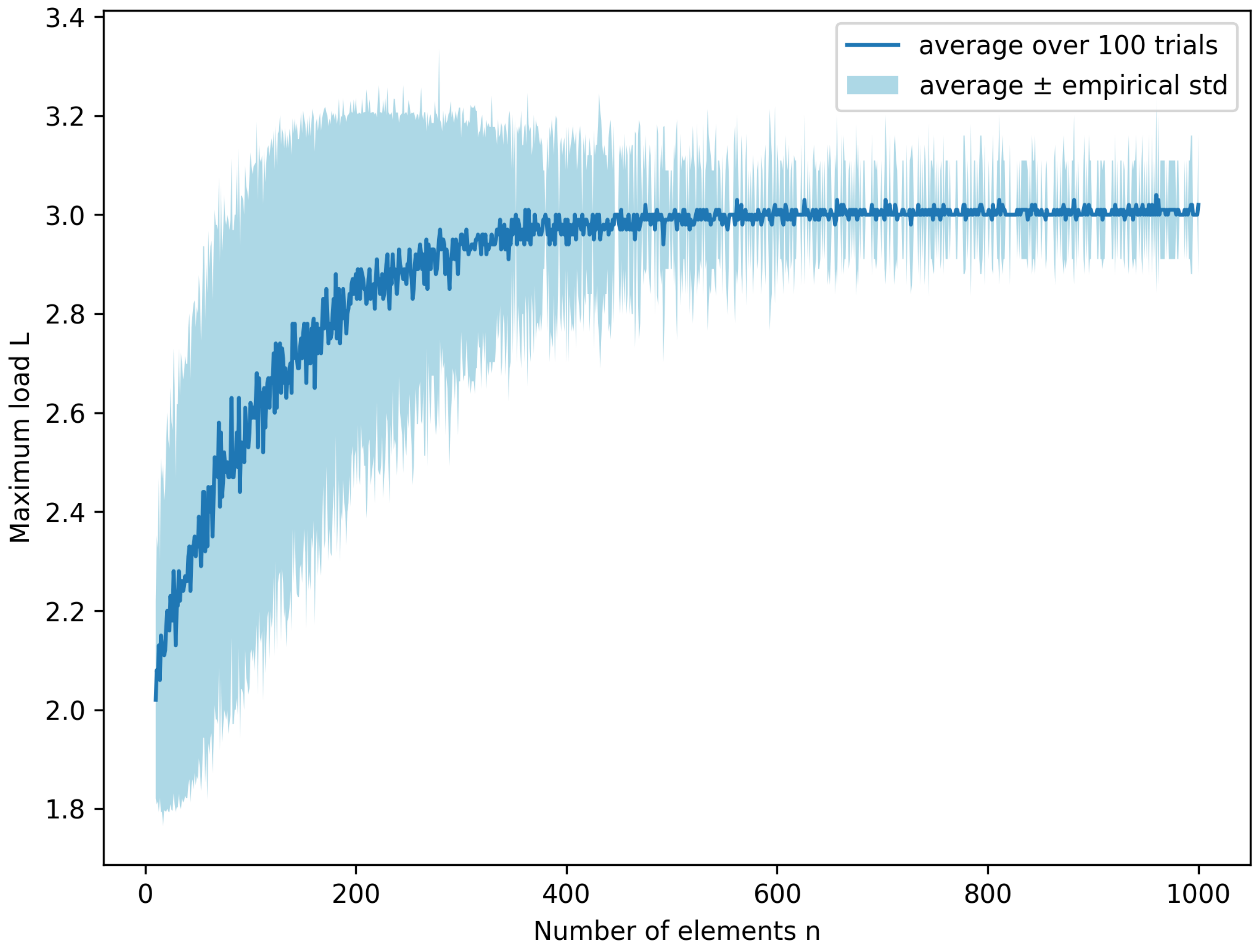
Load balancing (a twist)

"Now, every time you throw a ball, it selects two bins at random, and goes to the least full of the two. What is the maximum expected load?"

Load balancing (a twist)

"Now, every time you throw a ball, it selects two bins at random, and goes to the least full of the two. What is the maximum expected load?"

- $\Theta(1)$?
- $\Theta(\log \log n)$?
- $\Theta(\sqrt{\log n})$?
- Something else?


average over 100 trials
average ± empirical std
Maximum load L
Number of elements n
3.4
3.2
3.0
2.8
2.6
2.4
2.2
2.0
1.8
0
200
400
600
800
1000

Load balancing (a twist)

"Now, every time you throw a ball, it selects two bins at random, and goes to the least full of the two. What is the maximum expected load?"

Theorem. The expected maximum load now $\Theta(\log \log n)$.