Probability and Statistics

Tail Inequalities

谢润烁 Nanjing University, 2023 Fall

Roadmap

- Tail Inequalities: Part I
  - Markov Inequality
  - Chebyshev Inequality
- Concentration of Measure
- Tail Inequalities: Part II
  - Chernoff Bounds
- Application
  - Randomized Quick Sort
  - The Median Trick
  - Load Balancing Problem
- More General Bounds

Tail Inequalities: Part I

In Analysis of Algorithms

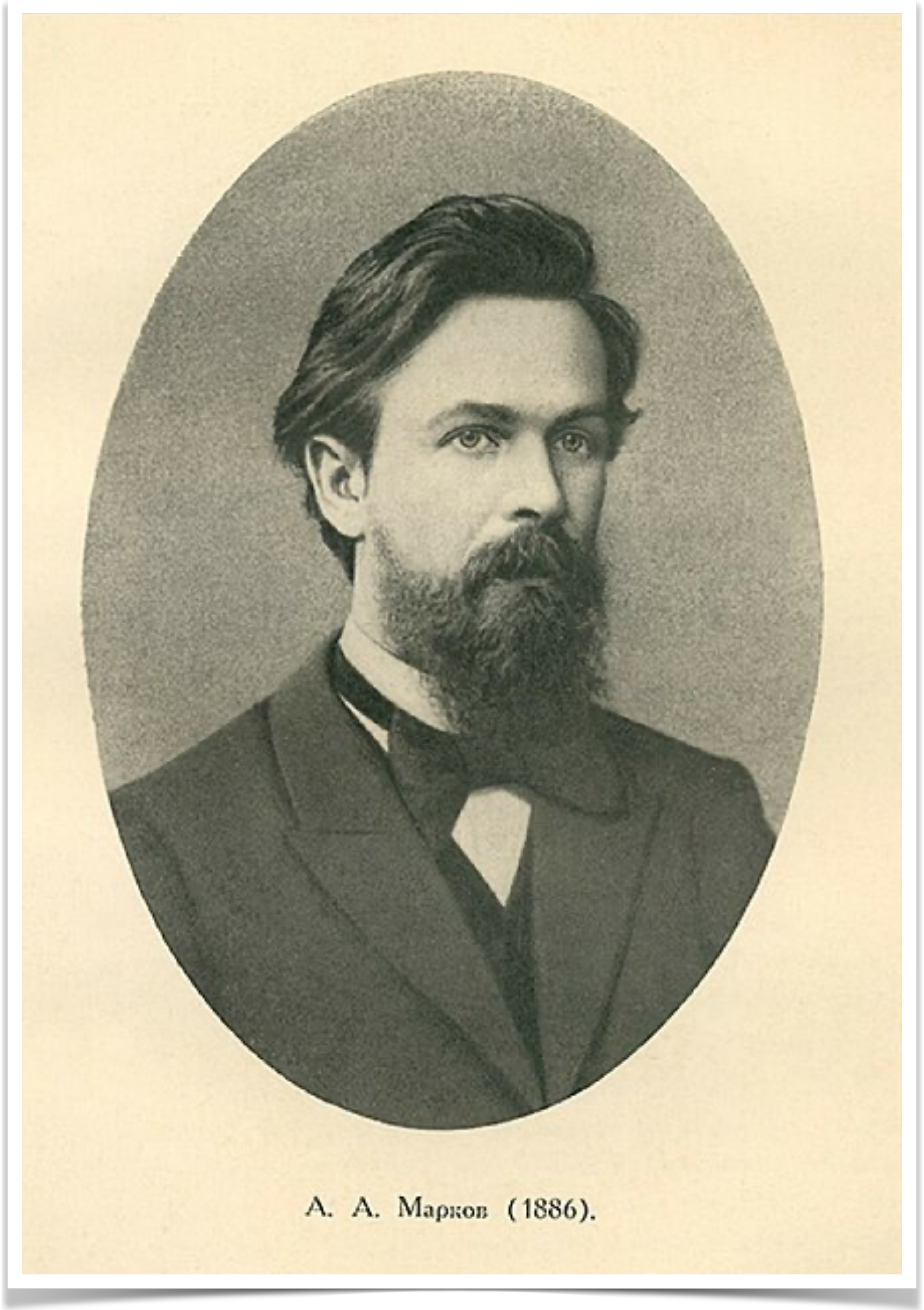
- Question: X is the running time of algorithm $\mathscr{A}$
  - Is it possible that $\Pr(X \geq \mathbb{E}X + t)$ very large?
- In analyzing the performance of a randomized algorithm, we often like to show that the behavior of the algorithm is good almost all the time
  - *i.e.* Establish high probability bounds on their run-time
  - *i.e.* Estimate the failure probability of algorithms

Markov's Inequality

- For any random variable $X \geq 0$ and $a > 0$

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

- Gives the best tail bound possible when all we know is
  - the expectation of the random variable and
  - the variable is nonnegative



1856 ~ 1922
Андре́й Андре́евич Ма́рков
Andrey Andreyevich Markov

Markov's Inequality

- For any random variable $X \geq 0$ and $a > 0$

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

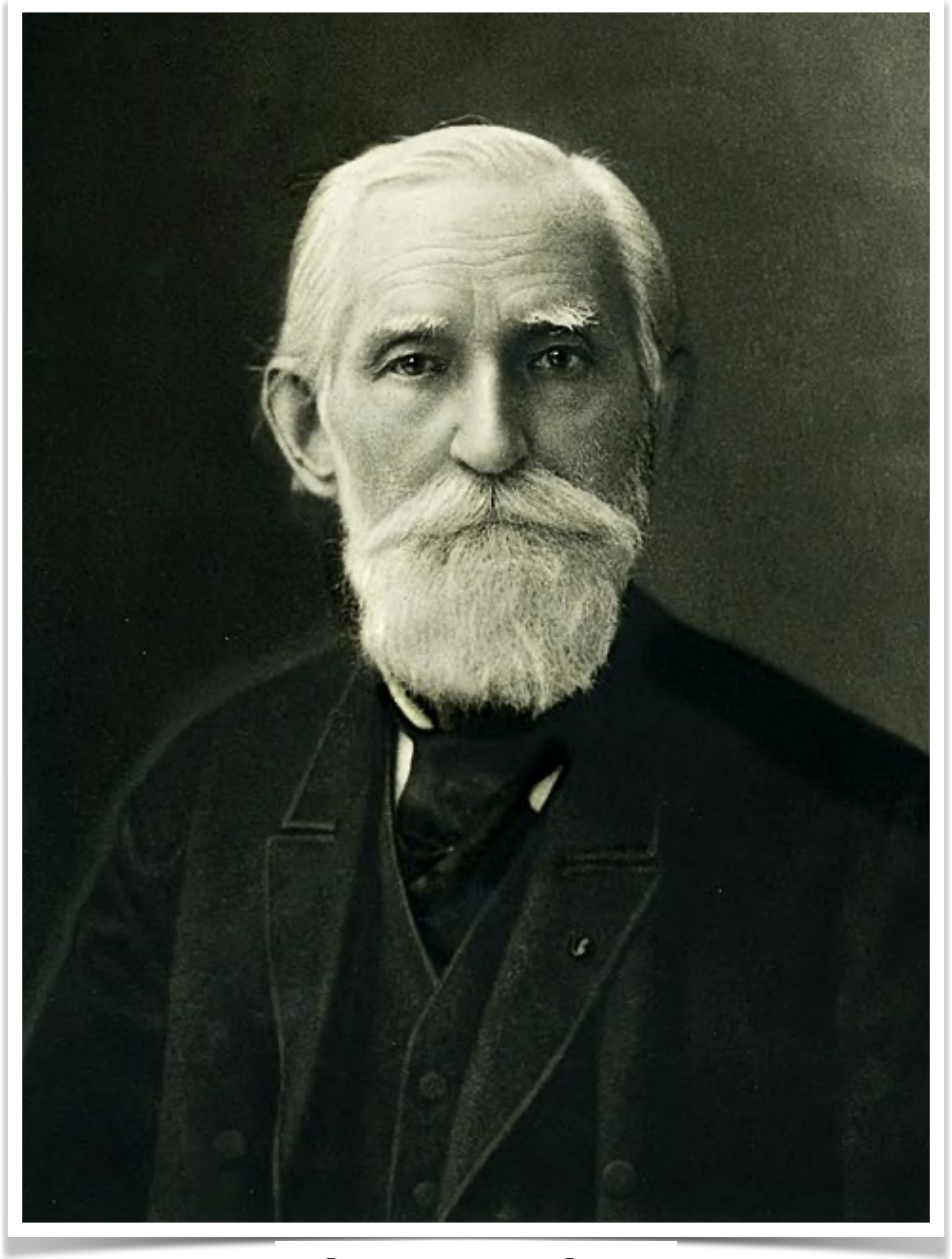
- Unfortunately, the Markov inequality is often too weak to yield useful results
- Can we leverage more information to gain a better bound?

Chebyshev’s Inequality

- For any $a > 0$

$$\Pr(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathrm{Var}[X]}{a^2}$$

- Proof: Apply Markov’s Inequality



1821 ~ 1894
Пафну́тий Льво́вич Чебышёв
Pafnuty Lvovich Chebyshev
[pɐfˈnutʲɪj ˈlʲvovʲɪtɕ tɕɪbɨˈʂof]

Chebyshev’s Inequality

- For any $a > 0$

$$\Pr(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathrm{Var}[X]}{a^2}$$

- Is it better than Markov’s Inequality?

Chebyshev's Inequality

- For any $a > 0$

$$\Pr(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathrm{Var}[X]}{a^2}$$

- Is it better than Markov's Inequality?
- Consider flipping coins $X \sim \mathrm{Bin}(n, \frac{1}{2})$
  - $\Pr(X \geq \frac{3}{4}n)$

Generalized Markov's Inequality

- For any $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ and $a > 0$

$$\Pr(f(X) \geq a) \leq \frac{\mathbb{E}[f(X)]}{a}$$

- Useful if $f(X)$ can "extract" useful information about X

Generalized Markov's Inequality

- For any $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ and $a > 0$

$$\Pr(f(X) \geq a) \leq \frac{\mathbb{E}[f(X)]}{a}$$

- Example
  - kth moment method:
    - $f(X) = \mathbb{E}[X^k]$ kth moment
    - $f(X) = \mathbb{E}[(X - \mathbb{E}X)^k]$ kth central moment
      - Chebyshev's inequality: $f(X) = \mathbb{E}[(X - \mathbb{E}X)^2] = \mathrm{Var}(X)$
  - Chernoff-Hoeffding bounds: $f(X) = \mathbb{E}[e^{\lambda X}]$

Example: Weierstrass's Approximation Theorem

- Let $f : [0,1] \to [0,1]$ be a continuous function. For any $\epsilon > 0$, there exists a polynomial such that

$$\sup_{x \in [0,1]} |p(x) - f(x)| \leq \epsilon$$

- For $x \in [0,1]$, let $Y_x \sim \mathrm{Bin}(n, x)$

$$p(x) = \mathbb{E}[f(\frac{Y_x}{n})]$$ is it a polynomial?

Example: Weierstrass's Approximation Theorem

- $|p(x) - f(x)| = \left| \mathbb{E}\left[f(\frac{Y_x}{n}) - f(x) \right] \right| \leq \mathbb{E}\left[\left| f(\frac{Y_x}{n}) - f(x) \right| \right]$
- Recall that
  - $f(x)$ is continuous in $[a, b] \Rightarrow f(x)$ is uniformly continuous in $[a, b]$
  - uniformly continuous $\Rightarrow \exists \delta > 0$ s.t. $|f(x) - f(y)| \leq \dfrac{\epsilon}{2}$ for all $|x - y| \leq \delta$

Example: Weierstrass's Approximation Theorem

$$|p(x) - f(x)| \leq \mathbb{E}\left[\left|f(\frac{Y_x}{n}) - f(x)\right|\right]$$

- Denote A as event $\left|\frac{Y_x}{n} - x\right| \leq \delta$
- Use conditional expectation

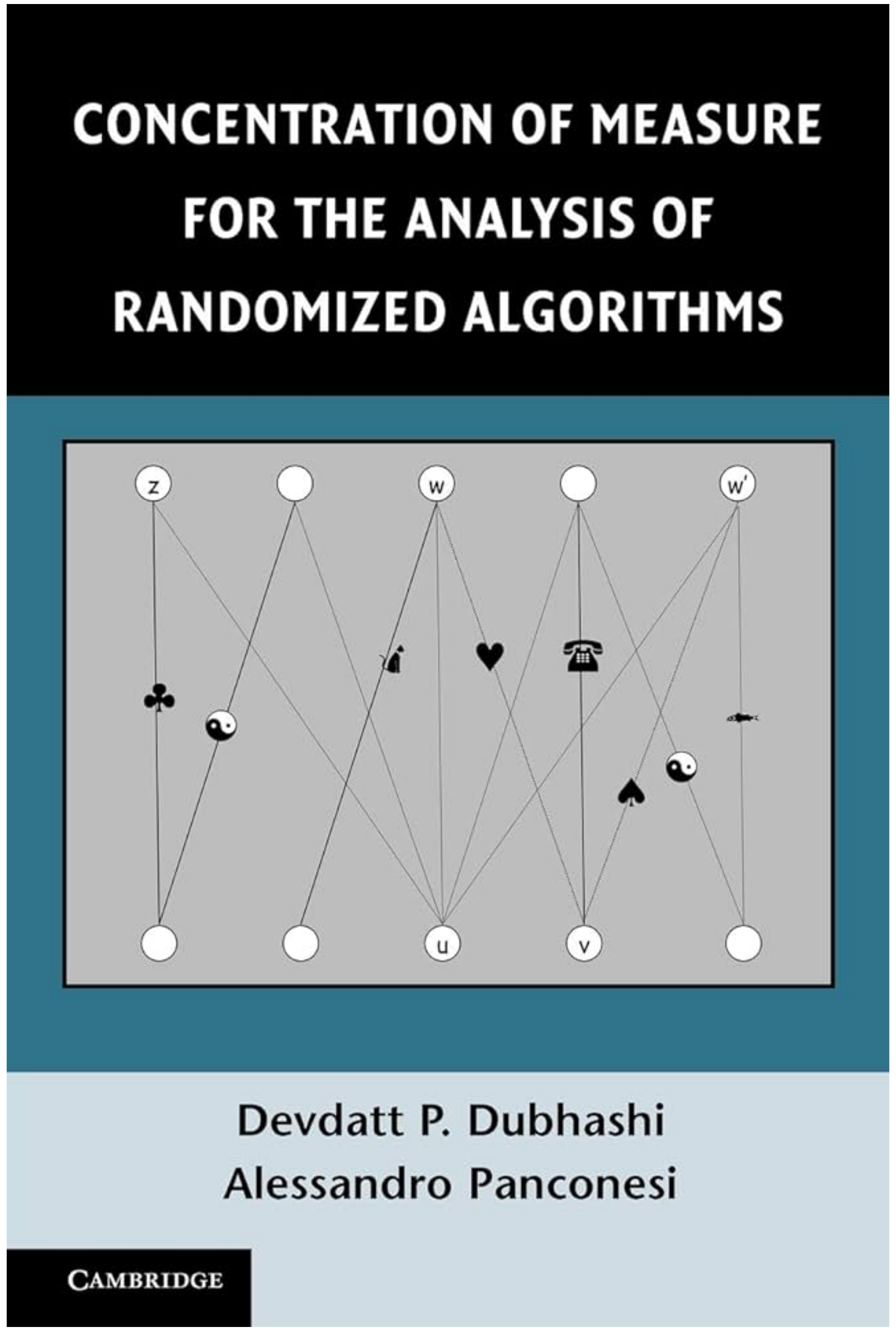
$$\mathbb{E}\left[\left|f(\frac{Y_x}{n}) - f(x)\right|\right] = \mathbb{E}\left[\left|f(\frac{Y_x}{n}) - f(x)\right| |A\right] \Pr(A) + \mathbb{E}\left[\left|f(\frac{Y_x}{n}) - f(x)\right| |A^c\right] \Pr(A^c)$$

$$\leq \frac{\epsilon}{2} + \frac{1}{4n\delta^2} \leq \epsilon \qquad \text{if } n \geq \frac{1}{2\epsilon\delta^2}$$

Concentration of Measure

What is *Concentration of Measure*?

- The phenomenon that a function of a **large number** of random variables tends to concentrate its values in a **relatively narrow range**.
- Views from different scale
    - Micro: random
    - Macro: regular



Law of Large Number

The weak version, aka Khinchin(Хи́нчин)'s Law

- $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables with mean μ and standard deviation σ
- For any constant $\varepsilon > 0$ we have

$$\lim_{n \to \infty} \Pr\left(\left| \frac{\sum_{i=1}^{n} X_i}{n} - \mu \right| > \epsilon \right) = 0$$

- You are able to prove this now

Central Limit Theorem

- $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables with mean μ and standard deviation σ

- For any real numbers a and b with $a < b$,

$$\lim_{n\to\infty} \Pr\left(a \leq \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \leq b\right) = \Phi(b) - \Phi(a)$$

- Where $\Phi(x)$ is the distribution function of standard normal distribution $N(0,1)$

Central Limit Theorem

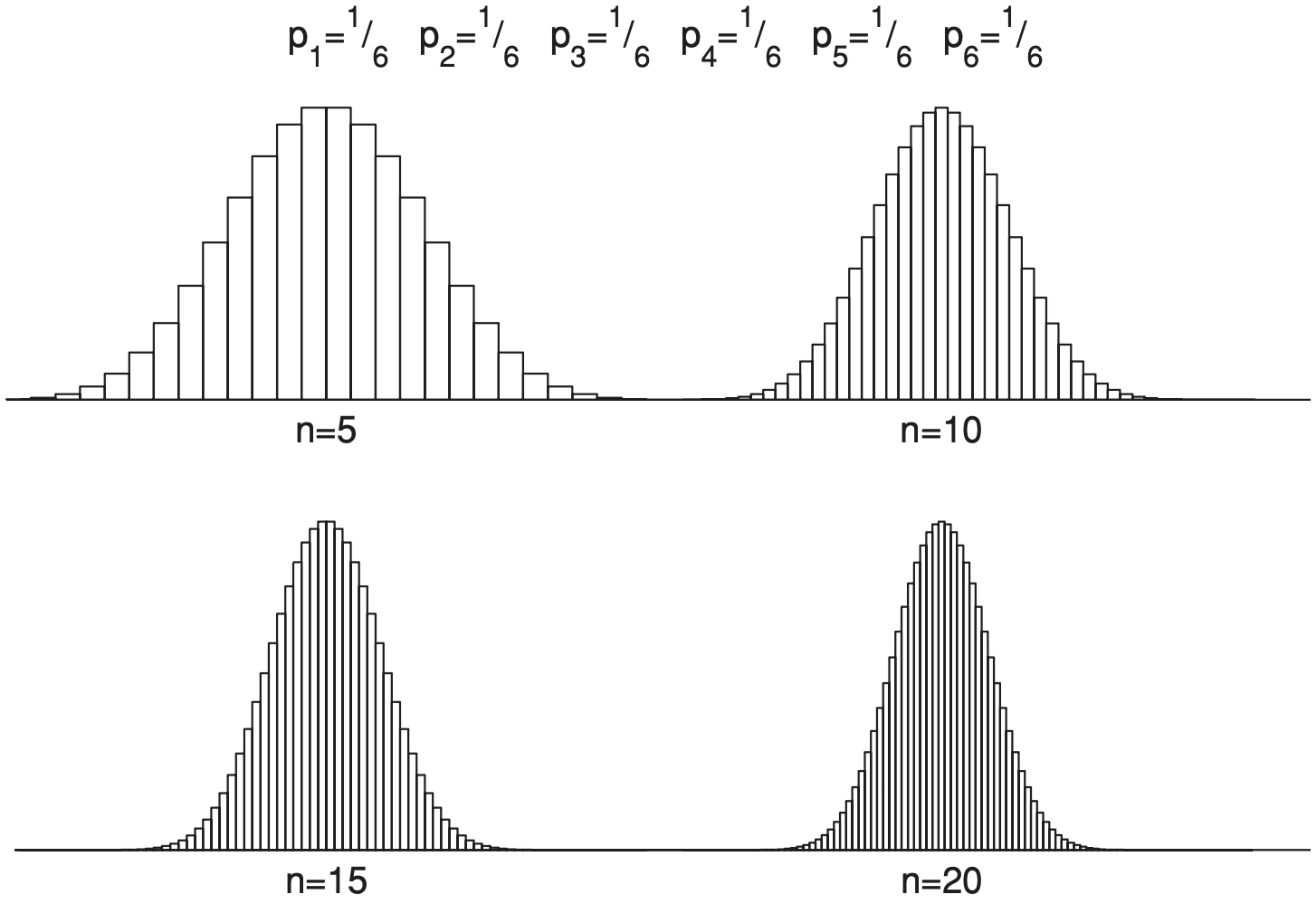


Fig. 5.6. Probability histogram for the unbiased die.

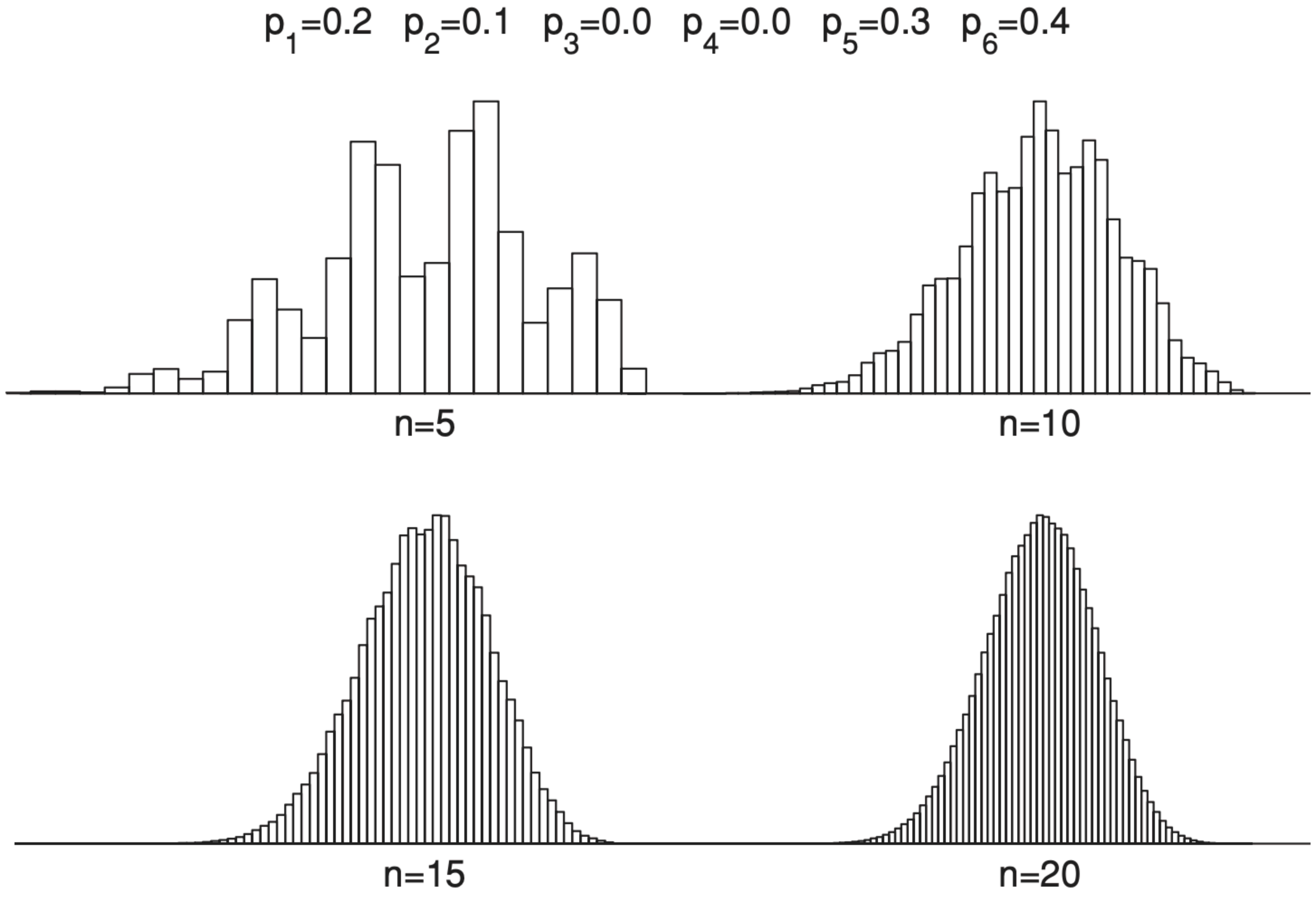


Fig. 5.7. Probability histogram for a biased die.

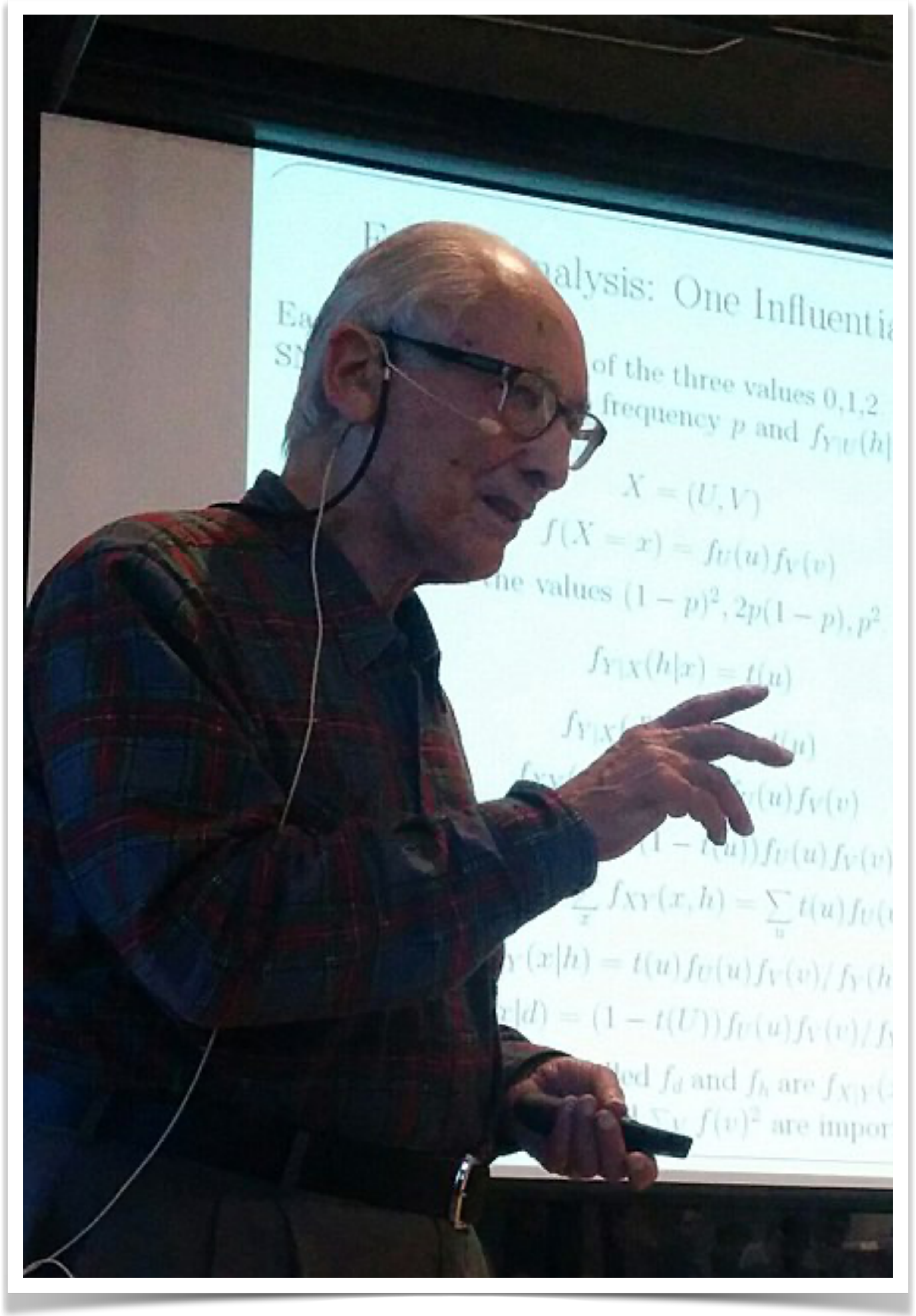
In Analysis of Algorithms

- These results (LLN, CLT) are **asymptotic** and **qualitative**
  - $n \to \infty$
- However, in the analysis of algorithms, we typically require **quantitative** estimates that are valid for finite (though large) values of n

Tail Inequalities: Part II

Chernoff and Hoeffding Bounds

- Extremely powerful in analysis of algorithms
- Giving exponentially decreasing bounds on the tail distribution
- Derived by applying Markov's inequality to the moment generating function of a random variable



1923 ~
Herman Chernoff

Moment Generating Function (MGF)

- Moment Generating Function

$$M_X(\lambda) = \mathbb{E}[e^{\lambda X}]$$

- In many cases, the function is well-defined in the neighborhood of zero
- Why *Moment Generating*?

$$\mathbb{E}[X_n] = M_X^{(n)}(0)$$

*: which is referred to as *Poisson Trial*

Chernoff Bounds

Tight Forms

Let $X = \sum_{i=1}^{n} X_i$ where $X_1, X_2, \ldots, X_n \in \{0,1\}$ are independent variables*

Let $\mu = \mathbb{E}[X]$

- for any $\delta \geq 0$,

$$\Pr[X \geq (1+\delta)\mu] \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}} \right)^{\mu} \text{(the Upper Tail)}$$

- for any $0 \leq \delta \leq 1$,

$$\Pr[X \leq (1-\delta)\mu] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}} \right)^{\mu} \text{(the Lower Tail)}$$

Proof Idea

On Upper Tail

- $\Pr[X \geq (1+\delta)\mu] = \Pr[e^{\lambda X} \geq e^{\lambda(1+\delta)\mu}] \leq \frac{\mathbb{E}\left[e^{\lambda X}\right]}{e^{\lambda(1+\delta)\mu}}$
- Find a λ to minimize $\frac{\mathbb{E}\left[e^{\lambda X}\right]}{e^{\lambda(1+\delta)\mu}}$

Chernoff Bounds

Useful Forms

- For any $0 < \delta < 1$,

$$\Pr[X \geq (1+\delta)\mu] \leq \exp\left(-\frac{\mu\delta^2}{3}\right)$$

$$\Pr[X \leq (1-\delta)\mu] \leq \exp\left(-\frac{\mu\delta^2}{2}\right)$$

- For $t \geq 2e\mu$,

$$\Pr[X \geq t] \leq 2^{-t}$$

Chernoff Bounds

- Compared to Markov's and Chebyshev's Inequalities
  - How is Chernoff Bounds' performance?
- Consider flipping coins $X \sim \mathrm{Bin}(n, \frac{1}{2})$ again
  - $\Pr(X \geq \frac{3}{4}n)$

Application

The Median Trick

- Suppose we want to estimate the value of m
- Let $\mathscr{A}$ be an algorithm that outputs $\widehat{Z}$ satisfying

$$\Pr[(1-\epsilon)m \leq \widehat{Z} \leq (1+\epsilon)m] \geq \frac{3}{4}$$

- How to improve our accuracy using $\mathscr{A}$?
- Let X be the median of $\widehat{Z}_1, \widehat{Z}_2, \ldots, \widehat{Z}_n$

$$\Pr[(1-\epsilon)m \leq X \leq (1+\epsilon)m] \geq ?$$

Randomized Quicksort

- We denote X as the running time of randomized quicksort, *i.e.*, #comparisons
  - You've learned in your DS course that
  - $\mathbb{E}(X) = \Theta(n \log n)$

Randomized Quicksort

智能软件与工程学院
School of Intelligent Software and Engineering

Randomized QuickSort

- Harmonic series
  - $H_n = \sum_{k=1}^{n} \frac{1}{k} \sim \ln n$
- Hence, $\mathbb{E}[X] < \sum_{i=1}^{n-1} \sum_{k=1}^{n} \frac{2}{k} < 2nH_n < 2n(1 + \ln n) = O(n \lg n)$
- Combined the fact that in the best case (balanced partition each time) randomized quick sort is $\Theta(n \lg n)$, the expected running time is $\Theta(n \lg n)$.
- In fact, runtime of `RndQuickSort` is $O(n \log n)$ with high probability!

Randomized Quicksort

- Now we can prove that the running time is $O(n \lg n)$ with high probability

$$\textit{i.e.} \lim_{n \to \infty} \Pr[X > O(n \lg n)] = 0$$

- Can we use the way we analyze the expected running time?

*$[n] = \{1,2,\ldots,n\}$

Load Balancing / Occupancy

Balls into Bins Model

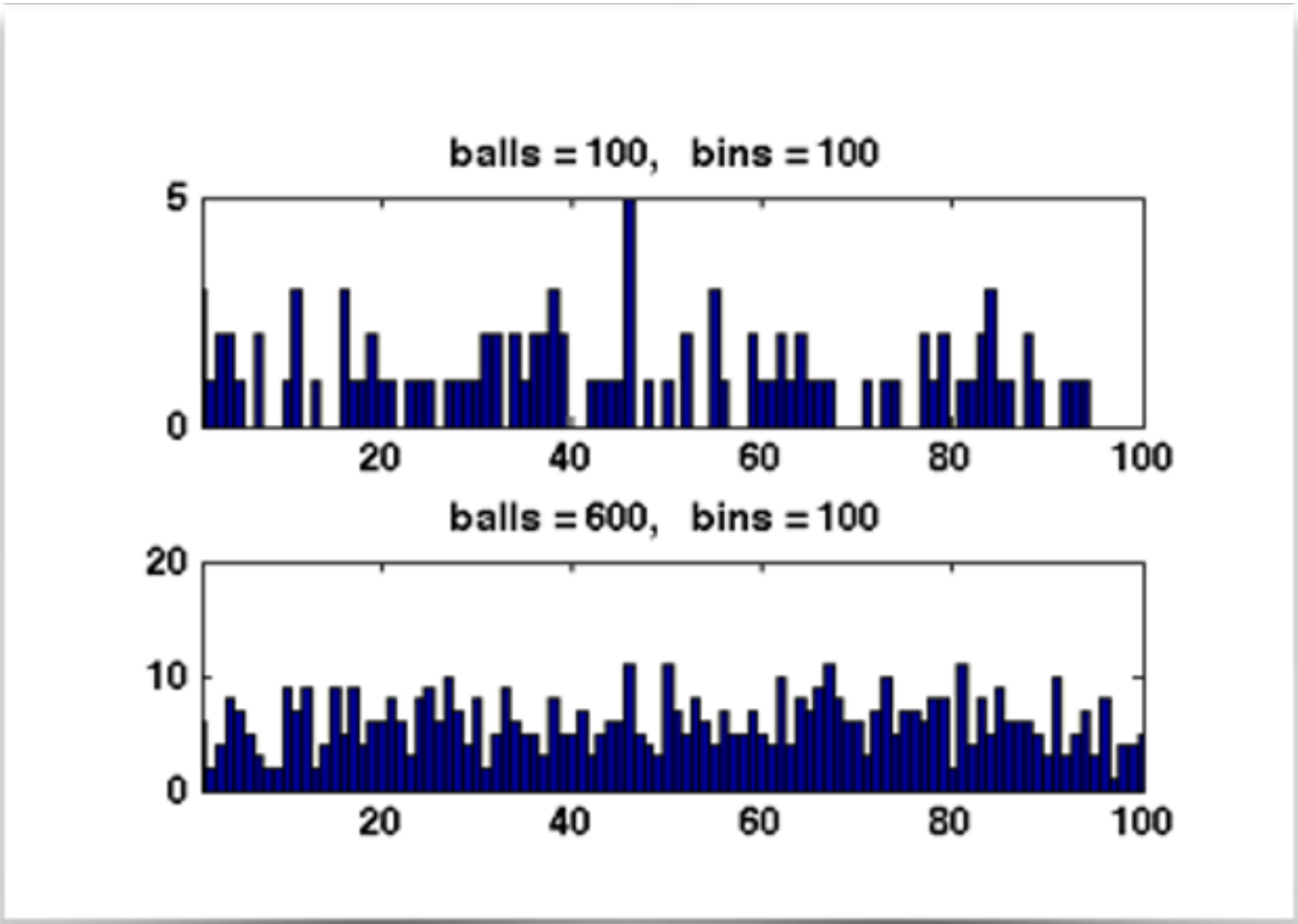
- We throw m balls into n bins uniformly and independently
- Y_i: number of balls, which is called the load, in the i-th bin

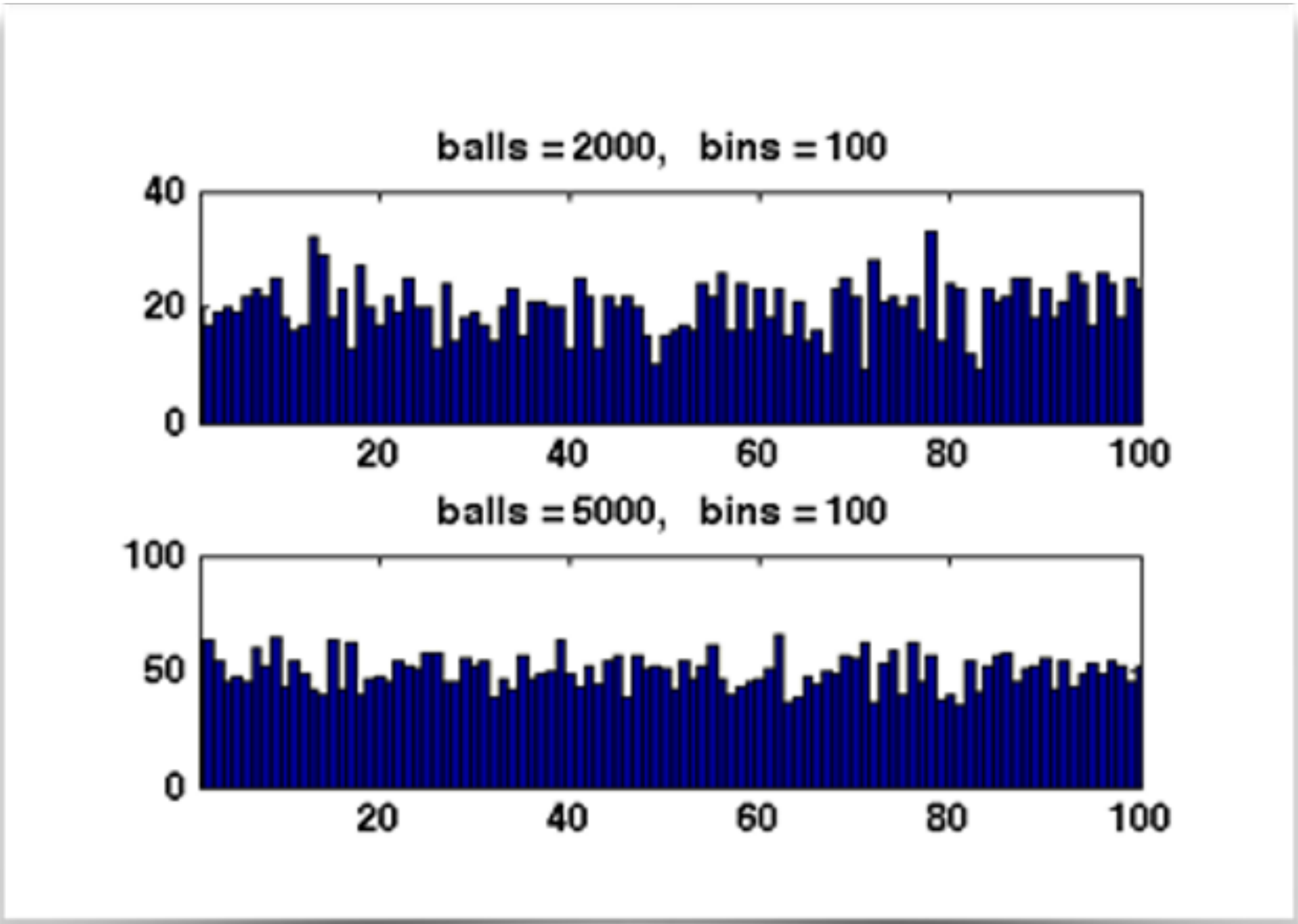
$$\mathbb{E}(Y_i) = \frac{m}{n}$$

- What is the maximum load of all bins?

Load Balancing / Occupancy

Balls into Bins Model





Load Balancing / Occupancy

Balls into Bins Model

- When $m = n$, the maximum load is

$$O\left(\frac{e \ln n}{\ln \ln n}\right) \text{ w.h.p.}$$

- When $m > n \ln n$, the maximum load is

$$O\left(\frac{m}{n}\right) \text{ w.h.p.}$$

More General Bounds

Chernoff-Hoeffding Bounds

- Let $X_1, ..., X_n$ be independent random variables with $\Pr(a_i \leq X_i \leq b_i) = 1$ for constants a_i and b_i. Then

$$\Pr\left(\left|X - \mu\right| \geq \varepsilon\right) \leq 2e^{\frac{-2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$$

- Where $X = \sum_{i=1}^{n} X_i$, $\mu = \mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{E}[X_i]$

The Method of Bounded Differences

- For independent $X_1, \ldots, X_n$, if n-variate function f satisfies the Lipschitz condition: for every $1 \leq i \leq n$ and all $x_1, \ldots, x_n$ and y_i

$$\left| f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n) \right| \leq c_i$$

- Then for any $\epsilon > 0$:

$$\Pr\left[\left| f(X_1, \ldots, X_n) - \mathbb{E}(f(X_1, \ldots, X_n) \right| \geq \epsilon \right] \leq 2e^{\frac{-2\epsilon^2}{\Sigma_{i=1}^{n} c_i}}$$