Introduction to Machine Learning CMU-10701

8. Stochastic Convergence

Barnabás Póczos

MACHINE LEARNING DEPARTMENT

Carnegie Mellon.
School of Computer Science

Motivation

What have we seen so far?

Several algorithms that seem to work fine on training datasets:

- Linear regression
- Naïve Bayes classifier
- Perceptron classifier
- Support Vector Machines for regression and classification

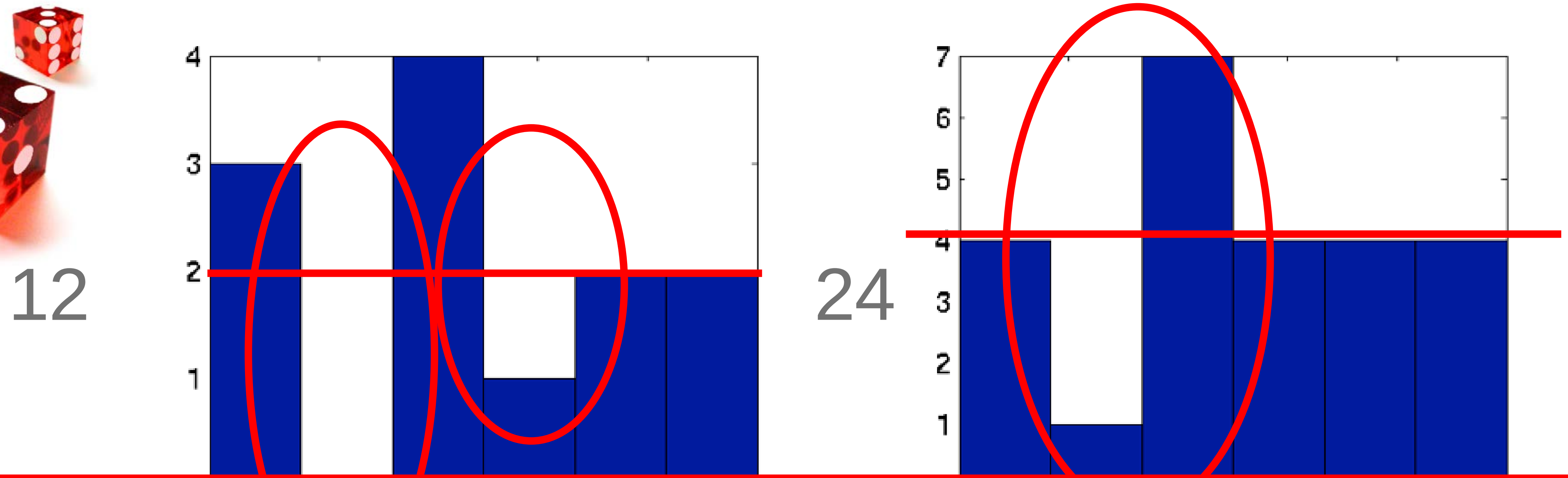
- ❑ How good are these algorithms on unknown test sets?
- ❑ How many training samples do we need to achieve small error?
- ❑ What is the smallest possible error we can achieve?

$\Rightarrow$ Learning Theory

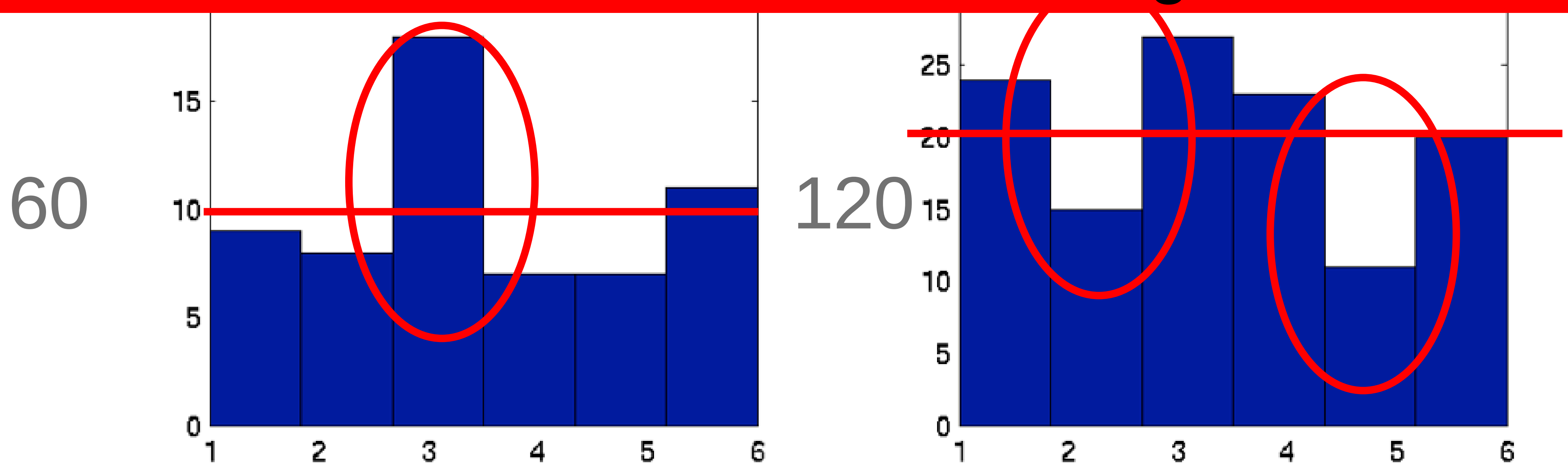
To answer these questions, we will need a few powerful tools

Basic Estimation Theory

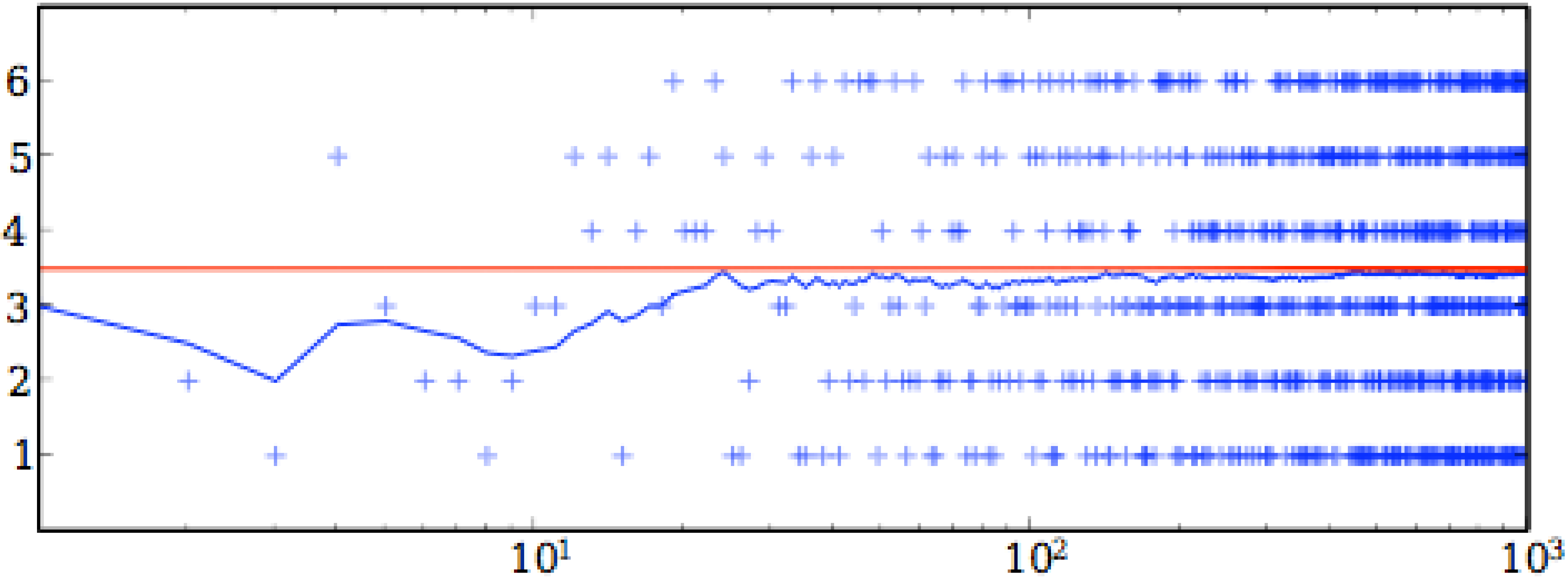
Rolling a Dice,
Estimation of parameters $\theta_1, \theta_2, \ldots, \theta_6$



Does the MLE estimation converge to the right value?
How fast does it converge?

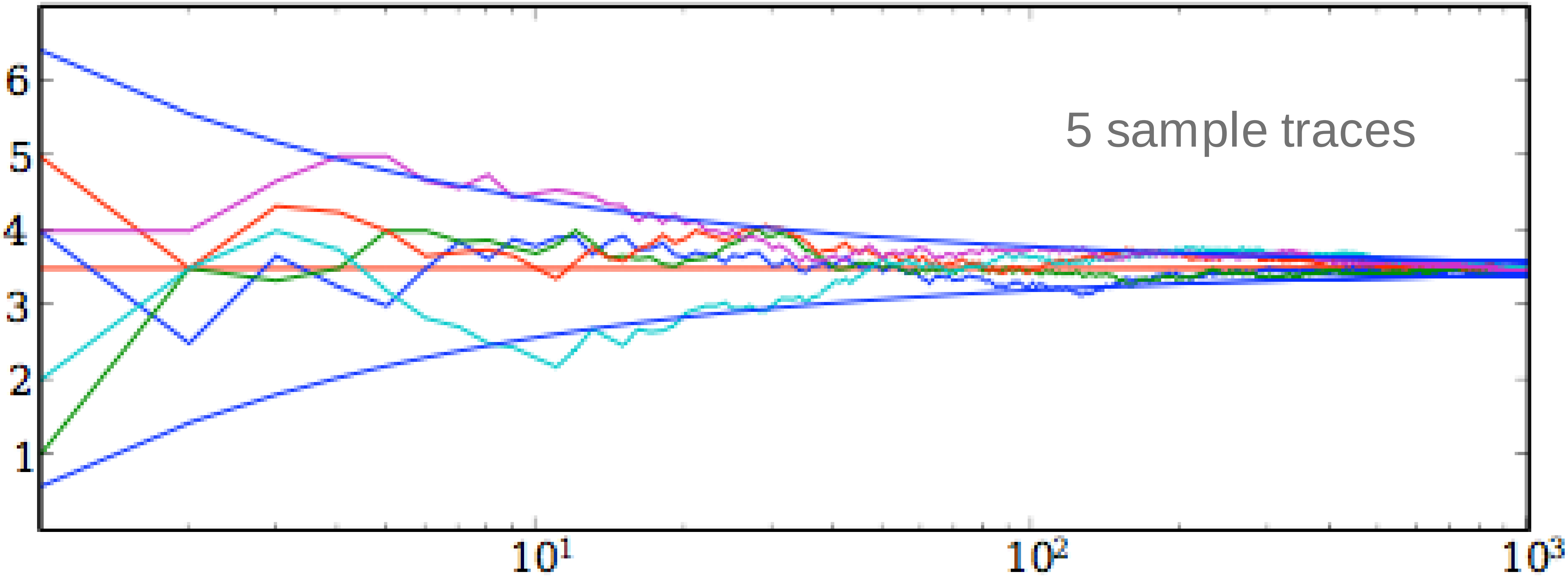


Rolling a Dice
Calculating the Empirical Average



Does the empirical average converge to the true mean?
How fast does it converge?

Rolling a Dice, Calculating the Empirical Average



How fast do they converge? $\mu \pm \sqrt{\mathrm{Var}(x)/n}$

Key Questions

- Do empirical averages converge?
- Does the MLE converge in the dice rolling problem?
- What do we mean on convergence?
- What is the rate of convergence?

I want to know the coin parameter $\theta \in [0,1]$ within $\varepsilon = 0.1$ error, with probability at least $1-\delta = 0.95$.
How many flips do I need?

Applications:

- drug testing (Does this drug modifies the average blood pressure?)
- user interface design (We will see later)

Outline

Theory:

- Stochastic Convergences:
  - Weak convergence = Convergence in distribution
  - Convergence in probability
  - Strong (almost surely)
  - Convergence in L_p norm

- Limit theorems:
  - Law of large numbers
  - Central limit theorem

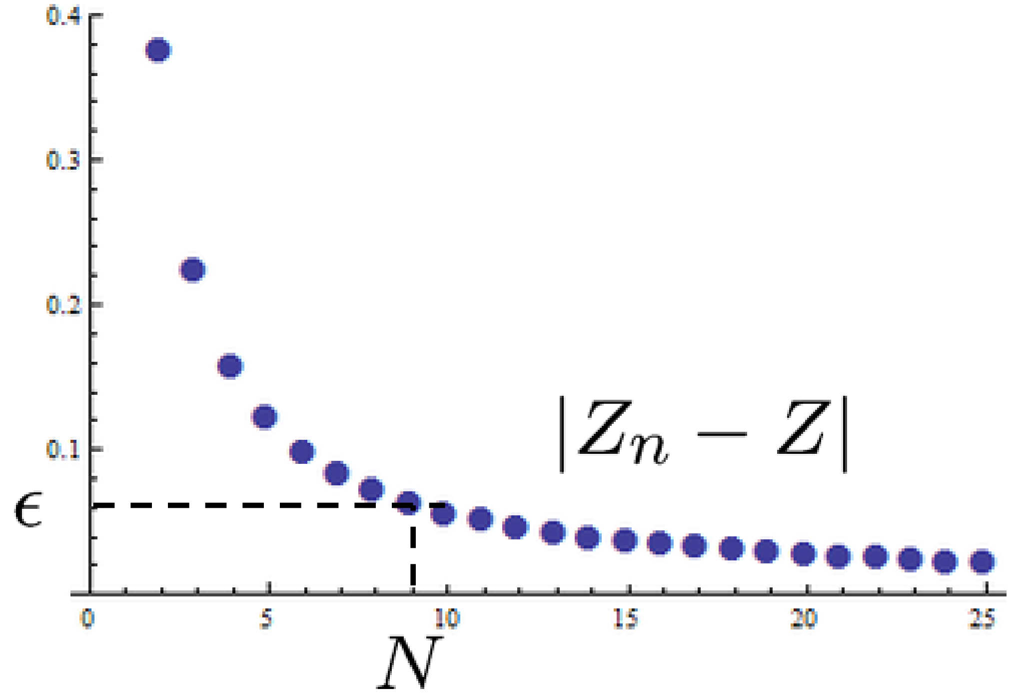
- Tail bounds:
  - Markov, Chebyshev

Stochastic convergence definitions and properties

Convergence of vectors

In $\mathbb{R}^n$ the $Z_n \to Z$ convergence definition is easy:

For each $\epsilon > 0$, there exists a $N > 0$ treshold number such that, for every $n > N$, we have $|Z_n - Z| < \epsilon$.



What do we mean on the convergence of random variables $Z_n \to Z$?

Convergence in Distribution = Convergence Weakly = Convergence in Law

Let $\{Z, Z_1, Z_2, \ldots\}$ be a sequence of random variables.
F_n and F are the cumulative distribution functions of Z_n and Z.

Notation:

$$Z_n \xrightarrow{d} Z, \quad Z_n \xrightarrow{\mathcal{D}} Z, \quad Z_n \xrightarrow{\mathcal{L}} Z, \quad Z_n \xrightarrow{d} \mathcal{L}_Z,$$
$$Z_n \rightsquigarrow Z, \quad Z_n \Rightarrow Z, \quad \mathcal{L}(Z_n) \to \mathcal{L}(Z), \quad F_n \xrightarrow{w} F$$

Definition:

$$\lim_{n\to\infty} F_n(z) = F(z), \ \forall z \in \mathbb{R} \text{ at which } F \text{ is continuous}$$

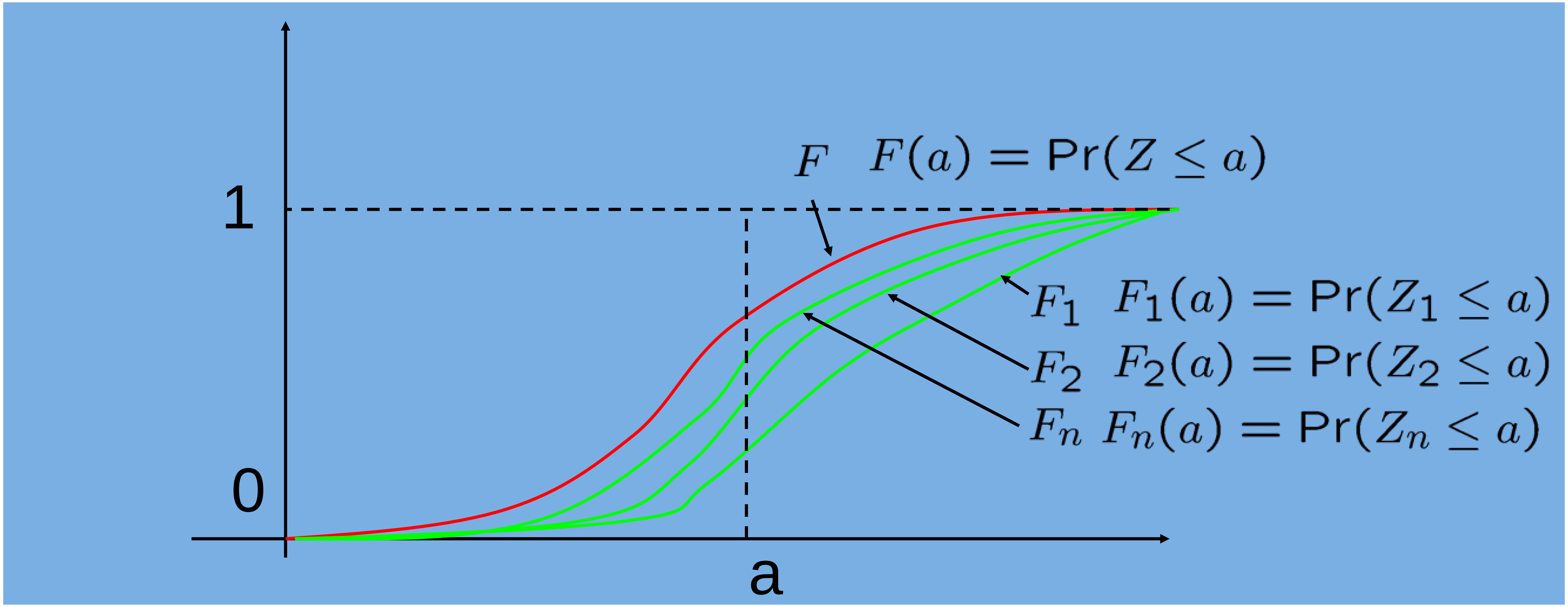
This is the "weakest" convergence.

Convergence in Distribution = Convergence Weakly = Convergence in Law

Only the distribution functions converge!
(NOT the values of the random variables)

$Z_n(\omega)$ can be very different of $Z(\omega)$

Random variable Z_n can be independent of random variable Z.

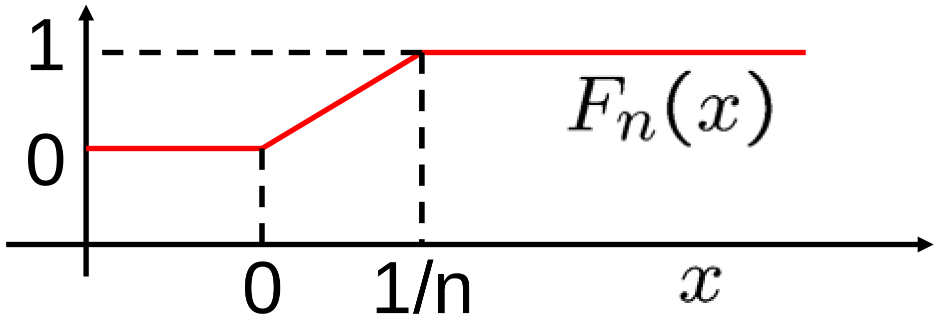


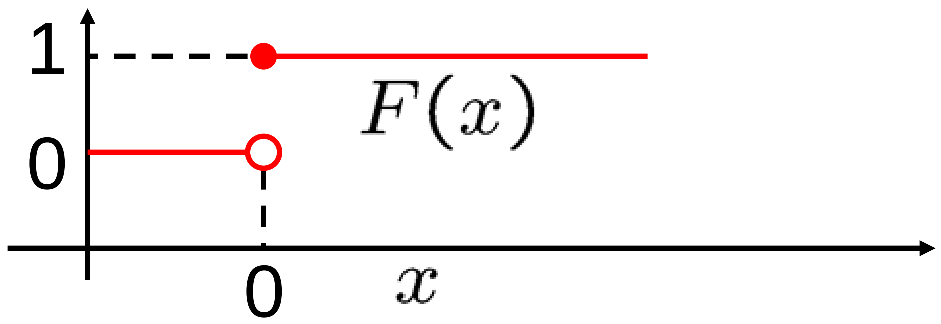
Convergence in Distribution = Convergence Weakly = Convergence in Law

Continuity is important!

Example: Let $Z_n \sim U[0, \frac{1}{n}]$. Then $Z_n \xrightarrow{d} 0$ degenerate variable.

Proof: $F_n(x) = 0$ when $x \leq 0$, and $F_n(x) = 1$ when $x \geq \frac{1}{n}$





The limit random variable is constant 0:

$F(0) = 1$, even though $F_n(0) = 0$ for all n.
$\Rightarrow$ the convergence of cdfs fails at $x = 0$ where F is discontinuous.

In this example the limit Z is discrete, not random (constant 0), although Z_n is a continuous random variable.

This image cannot currently be displayed.

Convergence in Distribution = Convergence Weakly = Convergence in Law

Properties

- For large n, $\Pr(Z_n \leq a) \approx \Pr(Z \leq a)$, $\forall a$ continuity point of F
  Z_n and Z can still be independent even if their distributions are the same!
- $\mathbb{E}[f(Z_n)] \to \mathbb{E}[f(Z)]$, if f is bounded continuous function.
- *Scheffe's theorem:*
  convergence of the probability density functions ⇒ convergence in distribution

$$p_{Z_n}(a) \xrightarrow{n\to\infty} p_Z(a), \text{ for all } a \Rightarrow Z_n \xrightarrow{d} Z.$$

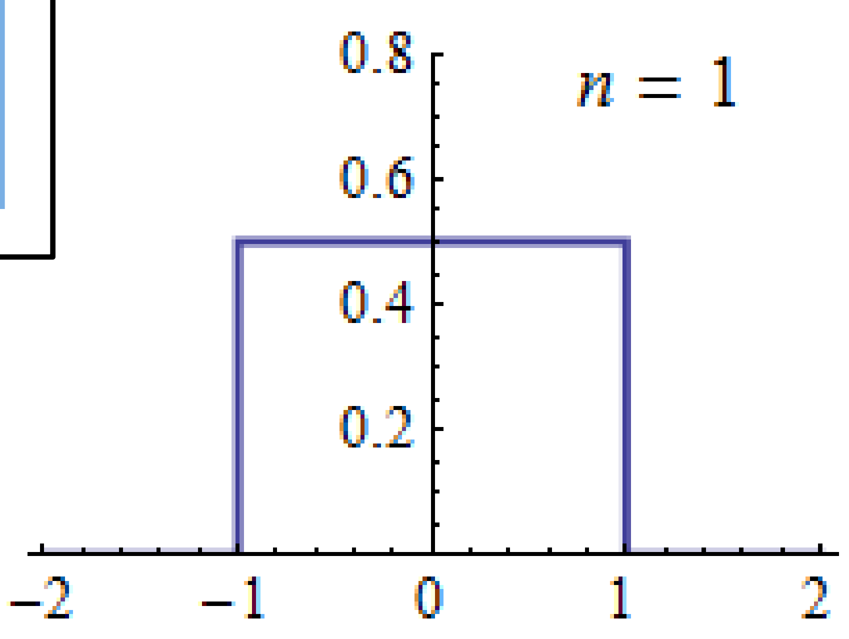
$$p_{Z_n}(a) \xrightarrow{n\to\infty} p_Z(a), \text{ for all } a \not\Leftarrow Z_n \xrightarrow{d} Z.$$

Example:
(Central Limit Theorem)

$$X_n \sim U[-1,1].$$

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i.$$

$$Z_n \xrightarrow{d} Z \sim \mathcal{N}(0,1)$$

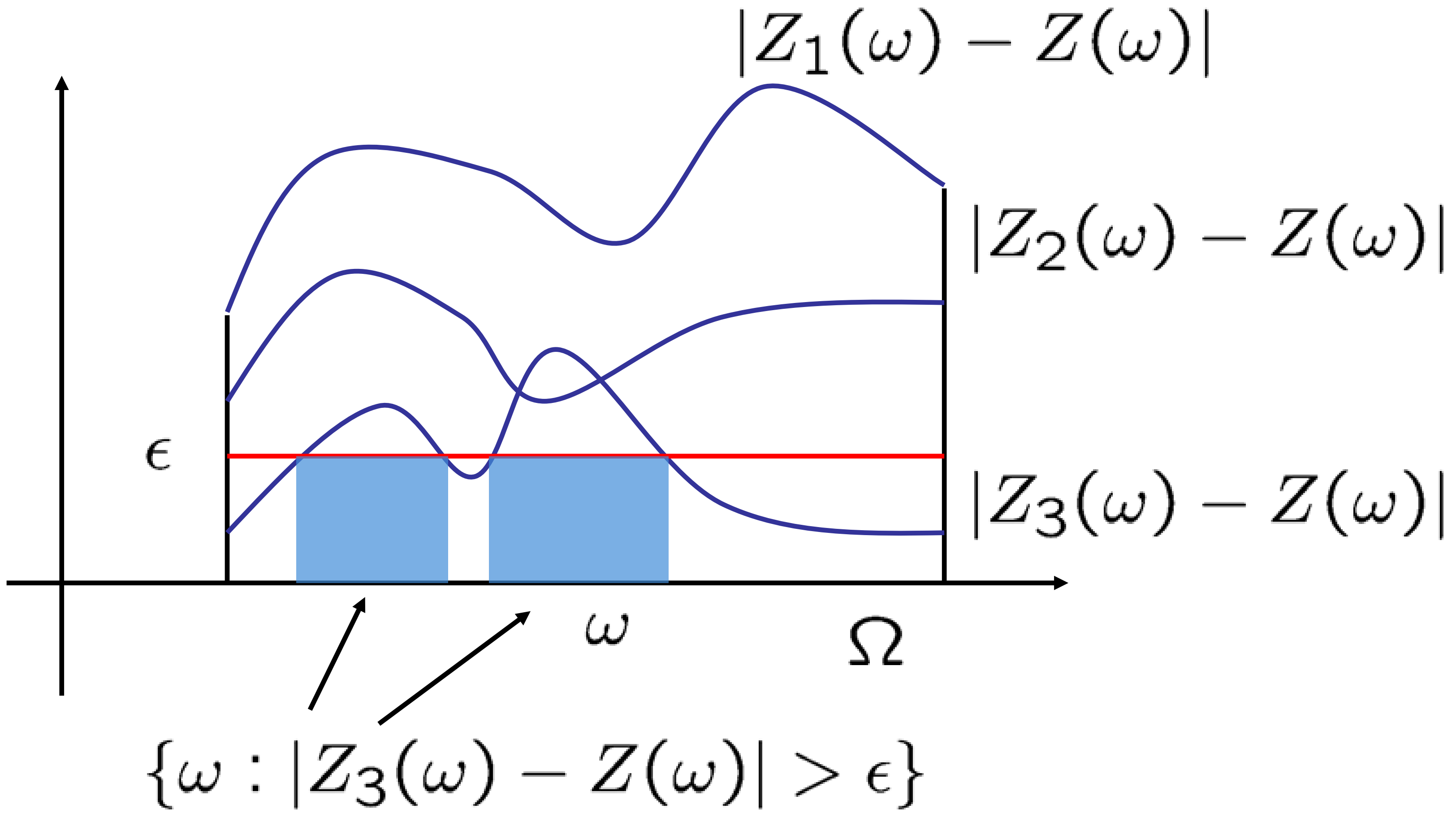


Convergence in Probability

Notation: $Z_n \xrightarrow{p} Z$

Definition: $\forall \varepsilon > 0 \ \lim_{n\to\infty} \Pr\left(|Z_n - Z| \geq \varepsilon\right) = 0.$

$\forall \varepsilon > 0 \ \lim_{n\to\infty} \Pr\left(|Z_n - Z| < \varepsilon\right) = 1.$

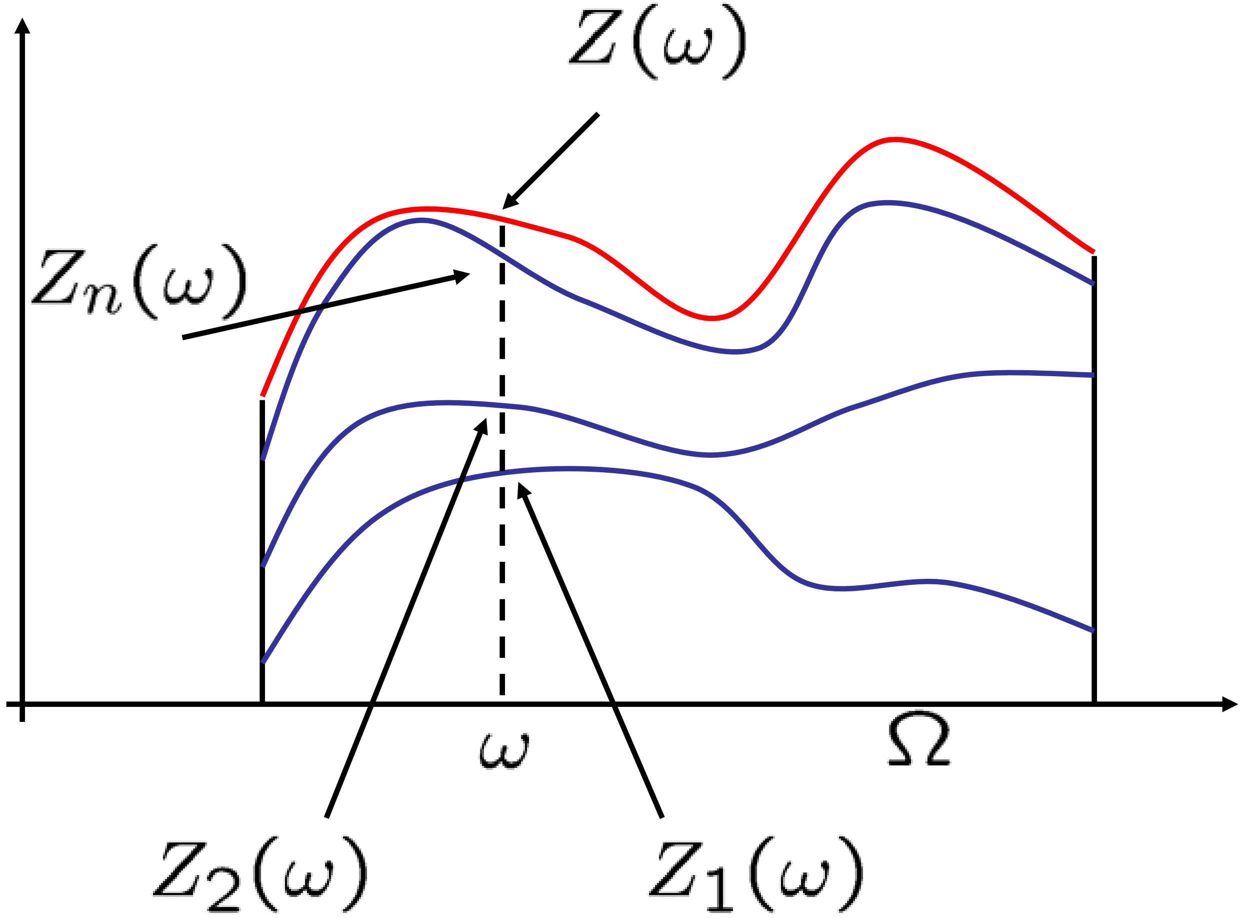


This indeed measures how far the values of $Z_n(\omega)$ and $Z(\omega)$ are from each other.

Almost Surely Convergence

Notation: $Z_n \xrightarrow{a.s.} Z \quad Z_n \to Z$ (w.p. 1)

Definition: $\Pr\left(\omega \in \Omega : \lim_{n\to\infty} Z_n(\omega) = Z(\omega)\right) = 1.$



Convergence in p-th mean, L_p norm

Notation: $Z_n \xrightarrow{L_p} Z$

Definition: $\lim_{n \to \infty} \mathbb{E}\left[|Z_n - Z|^p\right] = 0$

Properties:

$$Z_n \xrightarrow{a.s.} Z$$

$$\Downarrow$$

$$Z_n \xrightarrow{p} Z \Rightarrow Z_n \xrightarrow{d} Z$$

$$\Uparrow$$

$$Z_n \xrightarrow{L_p} Z$$

Counter Examples

$$Z_n \xrightarrow{d} Z \nRightarrow Z_n \xrightarrow{p} Z$$

$$Z_n \xrightarrow{p} Z \nRightarrow Z_n \xrightarrow{a.s.} Z$$

$$Z_n \xrightarrow{p} Z \nRightarrow Z_n \xrightarrow{L_p} Z$$

$$Z_n \xrightarrow{a.s.} Z \nRightarrow Z_n \xrightarrow{L_p} Z$$

$$Z_n \xrightarrow{L_p} Z \nRightarrow Z_n \xrightarrow{a.s.} Z$$

$$Z_n \xrightarrow{a.s.} Z$$

$$Z_n \xrightarrow{p} Z \Rightarrow Z_n \xrightarrow{d} Z$$

$$Z_n \xrightarrow{L_p} Z$$

$Z_n \xrightarrow{d} Z \Rightarrow \mathbb{E}[f(Z_n)] \to \mathbb{E}[f(Z)]$, if f is bounded continuous function.

$Z_n \xrightarrow{d} Z \nRightarrow \mathbb{E}[f(Z_n)] \to \mathbb{E}[f(Z)]$, if f is general function.

Further Readings on Stochastic convergence

- http://en.wikipedia.org/wiki/Convergence_of_random_variables
- **Patrick Billingsley**: Probability and Measure
- **Patrick Billingsley**: Convergence of Probability Measures

Finite sample tail bounds

Useful tools!

Gauss Markov inequality

If X is any nonnegative random variable and $a > 0$, then

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

Proof: Decompose the expectation

$$\begin{aligned}\Pr(X \geq a) &= \int_a^\infty p(x)dx \\ &\leq \int_a^\infty \frac{x}{a} p(x)dx = \frac{1}{a}\int_a^\infty xp(x)dx \\ &\leq \frac{1}{a}\int_0^\infty xp(x)dx = \frac{\mathbb{E}[X]}{a}\end{aligned}$$

Corollary: Chebyshev's inequality

Chebyshev inequality

If X is any nonnegative random variable and $a > 0$, then

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

Here Var(X) is the variance of X, defined as:

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

Proof:

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

Apply Gauss-Markov to $(X - \mathbb{E}[X])^2$ with a^2:

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

Generalizations of Chebyshev's inequality

Chebyshev: $\Pr(|X-\mu| \geq a) \leq \frac{\sigma^2}{a^2}$

where σ^2 is the variance and $\mu = \mathbb{E}[X]$ is the mean.

This is equivalent to this: $\Pr(-a \leq X-\mu \leq a) \geq 1-\frac{\sigma^2}{a^2}$

Symmetric two-sided case (X is symmetric distribution**)**

$$\Pr(k_1 < X < k_2) \geq 1 - \frac{4\sigma^2}{(k_2-k_1)^2}$$

Asymmetric two-sided case (X is asymmetric distribution**)**

$$\Pr(k_1 < X < k_2) \geq \frac{4[(\mu-k_1)(k_2-\mu)-\sigma^2]}{(k_2-k_1)^2}$$

There are lots of other generalizations, for example multivariate *X*.

Higher moments?

Markov: $\Pr(|X-\mu| \geq a) \leq \frac{\mathbb{E}[|X-\mu|]}{a}$

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

Higher moments: $\Pr(|X-\mu| \geq a) \leq \frac{\mathbb{E}(|X-\mu|^n)}{a^n}$

where n ≥ 1

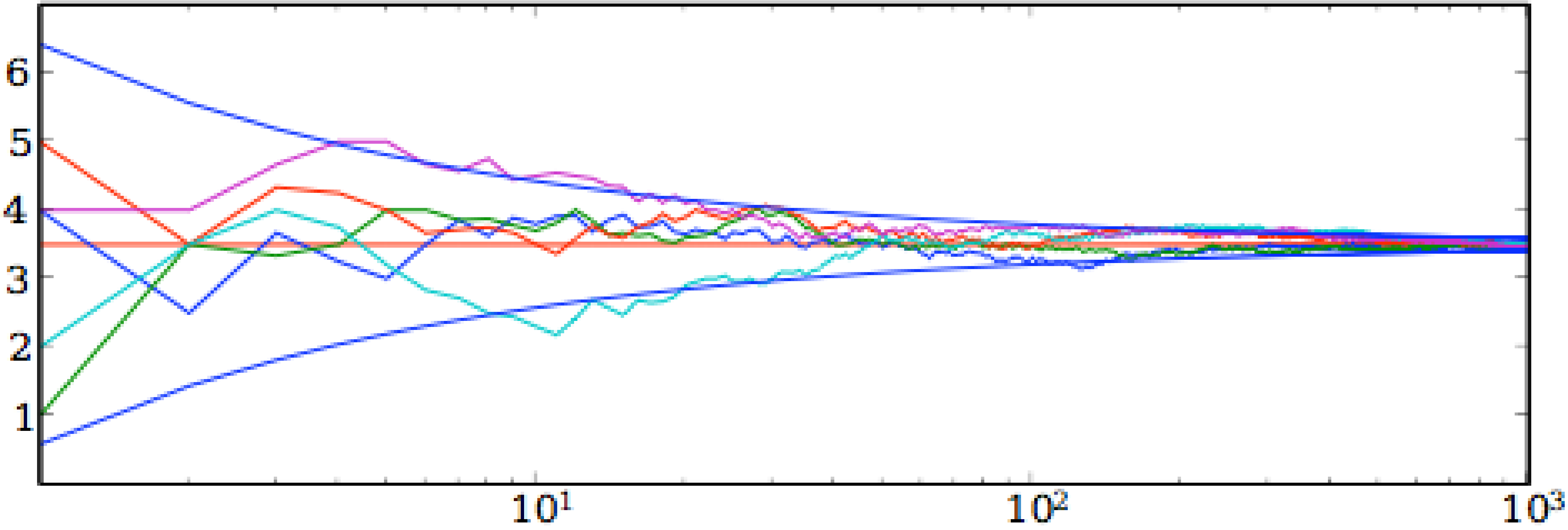
Other functions instead of polynomials?

Exp function: $\Pr(X \geq a) \leq e^{-ta}\mathbb{E}(e^{tX})$ where $a, t, X \geq 0$

Proof: $\Pr(X \geq a) = \Pr(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}}$ (Markov ineq.)

Law of Large Numbers

Do empirical averages converge?



Chebyshev's inequality is good enough to study the question:

Do the empirical averages converge to the true mean?

Answer: Yes, they do. (Law of large numbers)

Law of Large Numbers

$X_1, \ldots, X_n$ i.i.d. random variables with mean $\mu = \mathbb{E}[X_i]$

Empiricial average: $\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$

Weak Law of Large Numbers: $\widehat{\mu}_n \xrightarrow{p} \mu$

$$\forall \varepsilon > 0 \lim_{n \to \infty} \Pr\left(|\widehat{\mu}_n - \mu| \geq \varepsilon\right) = 0.$$

Strong Law of Large Numbers: $\widehat{\mu}_n \xrightarrow{a.s.} \mu$

$$\Pr\left(\omega \in \Omega : \lim_{n \to \infty} \widehat{\mu}_n(\omega) = \mu\right) = 1.$$

Weak Law of Large Numbers

Proof I:

$X_1, \ldots, X_n$ i.i.d., $\mu = \mathbb{E}[X_i]$ $\quad \widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

Assume finite variance. (Not very important) $\text{Var}(X_i) = \sigma^2$, (for all i)

$$\text{Var}(\widehat{\mu}_n) = \text{Var}(\tfrac{1}{n}(X_1 + \cdots + X_n)) = \frac{1}{n^2}\text{Var}(X_1 + \cdots + X_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

$\mathbb{E}[\widehat{\mu}_n] = \mu.$

Using Chebyshev's inequality on $\widehat{\mu}_n$ results in $\Pr(|\widehat{\mu}_n - \mu| \geq \varepsilon) \leq \dfrac{\sigma^2}{n\varepsilon^2}.$

Therefore,

$$\Pr(|\widehat{\mu}_n - \mu| < \varepsilon) = 1 - \Pr(|\widehat{\mu}_n - \mu| \geq \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}.$$

As *n* approaches infinity, this expression approaches 1.

$$\Rightarrow \widehat{\mu}_n \xrightarrow{P} \mu \qquad \text{for} \qquad n \to \infty.$$

What we have learned today

Theory:

- Stochastic Convergences:
  - Weak convergence = Convergence in distribution
  - Convergence in probability
  - Strong (almost surely)
  - Convergence in L_p norm

- Limit theorems:
  - Law of large numbers
  - Central limit theorem

- Tail bounds:
  - Markov, Chebyshev

Thanks for your attention ☺