Introduction to Machine Learning CMU-10701

2. MLE, MAP

What happened last time?

Barnabás Póczos & Aarti Singh

2014 Spring

Administration

- Piazza: … Please use it!
- Blackboard is ready
- Self assessment questions?
- Slides are online
- HW questions next week
- Feedback is important!
- Recitation: This Wednesday at 6pm (prob theory)

Independence

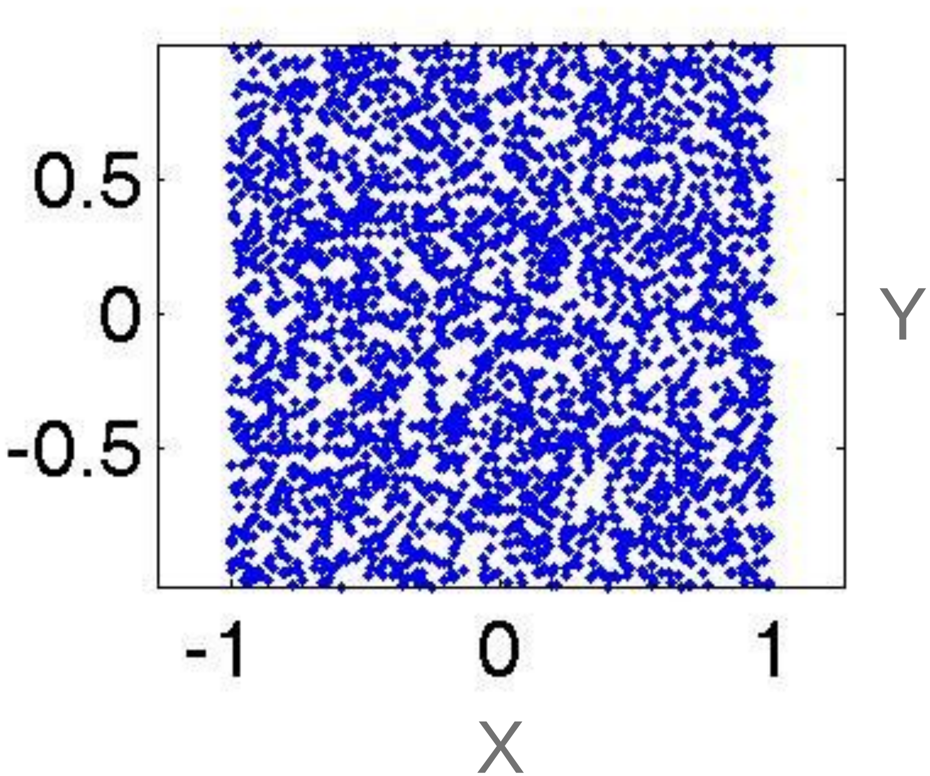
Independent random variables:

$$P(X,Y) = P(X)P(Y)$$

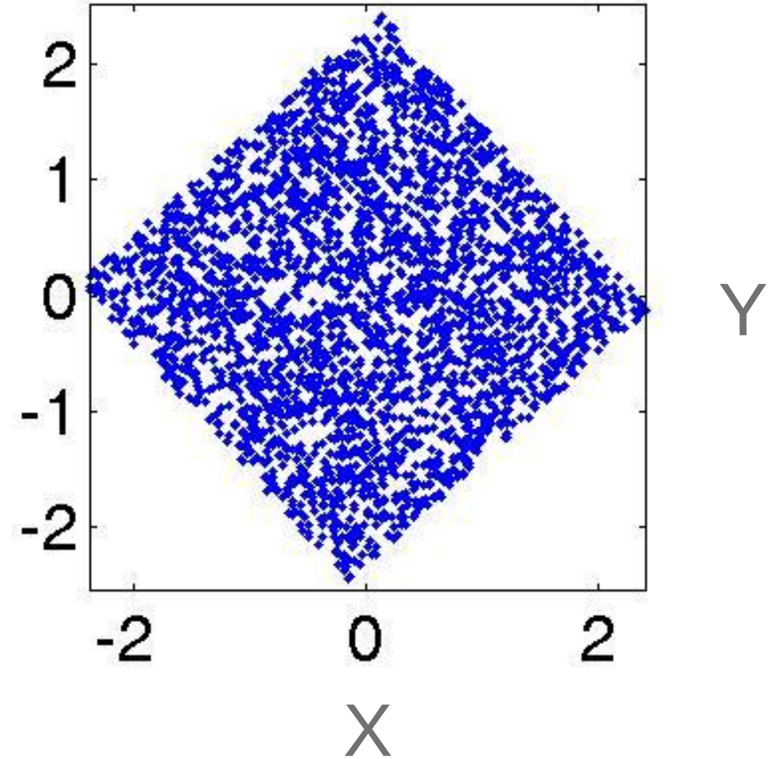
$$P(X|Y) = P(X)$$

Y and X don't contain information about each other.
Observing Y doesn't help predicting X.

Dependent / Independent



Independent X,Y



Dependent X,Y

Conditionally Independent

Conditionally independent:

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

Knowing Z makes X and Y independent

Examples:

Dependent: show size and reading skills
Conditionally independent: show size and reading skills given age

Our first machine learning problem:

Parameter estimation: MLE, MAP

MLE for Bernoulli distribution

Data, *D* =

$$D = \{X_i\}_{i=1}^{n}, \;\; X_i \in \{\mathrm{H}, \mathrm{T}\}$$

P(Heads) = θ*, P(Tails) = 1-*θ

The estimated probability is: **3/5** "Frequency of heads"

MLE: Choose θ that maximizes the probability of observed data

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

Independent draws

$$= \arg\max_{\theta} \prod_{i:X_i=H} \theta \prod_{i:X_i=T} (1-\theta)$$

Identically distributed

$$= \arg\max_{\theta} \; \underbrace{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}_{J(\theta)}$$

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

How good is this estimator?

I want to know the coin parameter $\theta \in [0,1]$ within $\varepsilon = 0.1$

error, with probability at least $1-\delta = 0.95$.

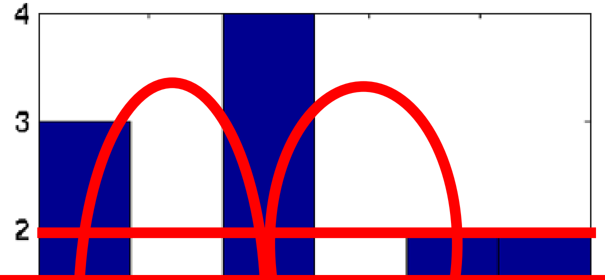
How many flips do I need?

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

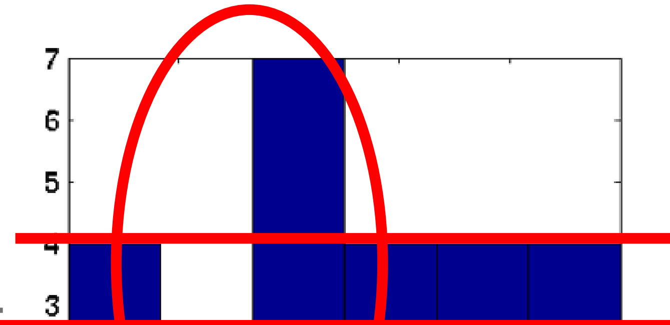
$$\Pr(|\widehat{\theta}_n - \theta| > \varepsilon) \leq \delta, \; n = ???$$

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

12



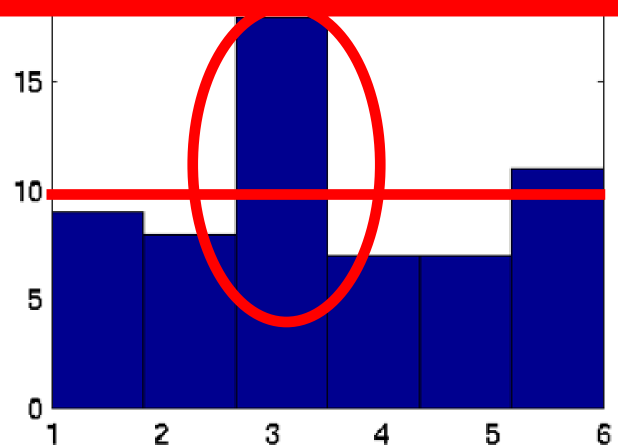
24



Does the MLE estimation (relative frequancies) converge to the right value?

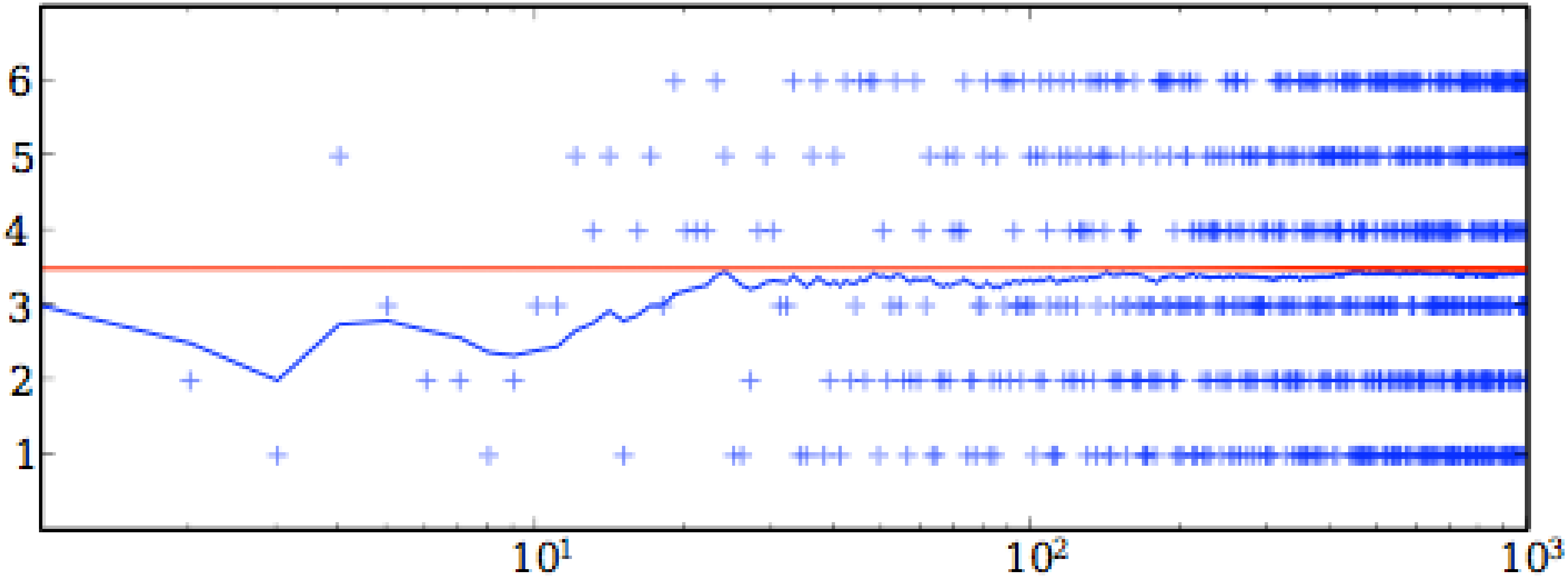
How fast does it converge?

60



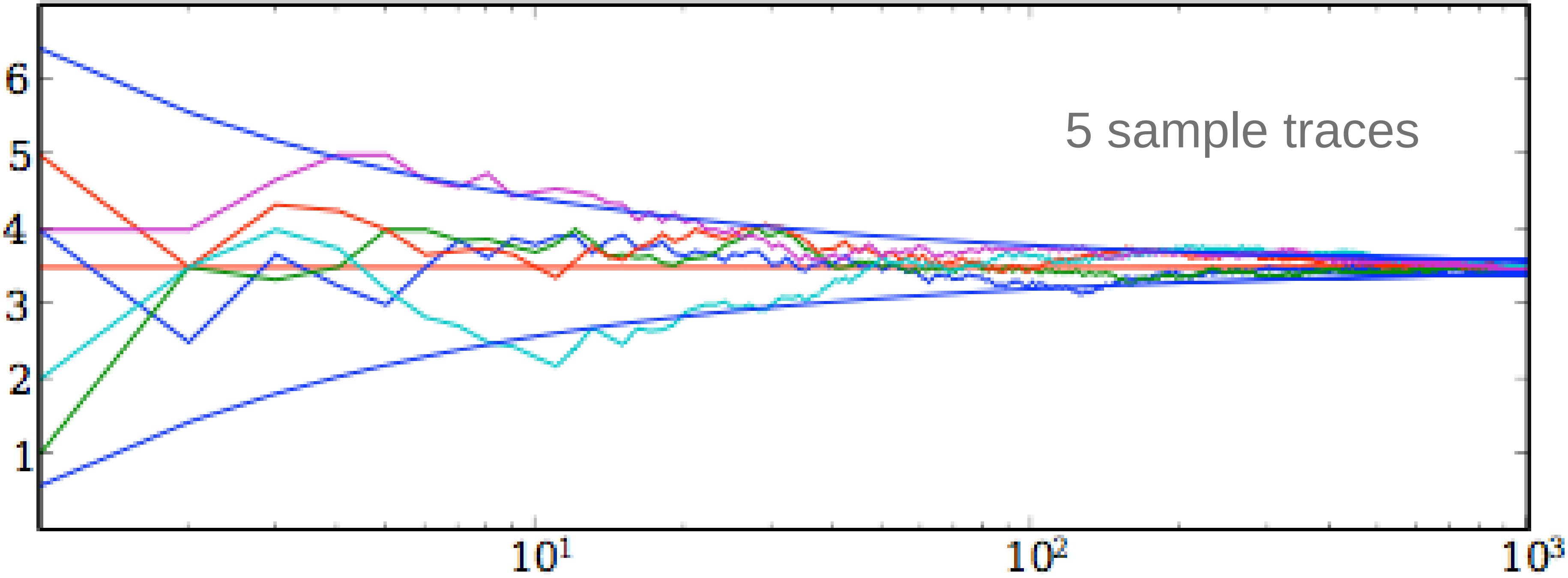
120

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 to the true mean?

$$\theta \pm \sqrt{Var(X)/n}$$

Hoeffding's inequality (1963)

$$\left.\begin{array}{r} X_1, ..., X_n \text{ independent} \\ X_i \in [a_i, b_i] \\ \varepsilon > 0 \end{array}\right\} \Rightarrow$$

$$\Rightarrow \mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)| > \varepsilon) \leq 2\exp\left(\frac{-2n\varepsilon^2}{\frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

It only contains the range of the variables, but not the variances.

"Convergence rate" for LLN from Hoeffding

From Hoeffding: Let $c^2 = \frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2$

$$\Rightarrow \Pr(|\widehat{\theta}_n - \theta| > \varepsilon) \leq 2\exp\left(\frac{-2n\varepsilon^2}{c^2}\right)$$

$$\delta = 2\exp\left(\frac{-2n\varepsilon^2}{c^2}\right)$$

$$\log\frac{\delta}{2} = \frac{-2n\varepsilon^2}{c^2}$$

$$\frac{c^2}{2n}\log\frac{2}{\delta} = \varepsilon^2$$

$$\varepsilon = c\sqrt{\frac{\log 2 - \log\delta}{2n}}$$

$\Rightarrow \left|\widehat{\theta}_n - \theta\right| < \varepsilon = c\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$ with prob. at least $(1-\delta)$

Convergence rate

Introduction to Machine Learning CMU-10701

Stochastic Convergence and Tail Bounds

Barnabás Póczos

MACHINE LEARNING DEPARTMENT
Carnegie Mellon.
School of Computer Science