CSE 312

Foundations of Computing II

Lecture 24: Biased Estimation

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Stefano Tessaro
tessaro@cs.washington.edu

Parameter Estimation – Workflow

Distribution $\mathbb{P}(x|\theta)$ → Independent samples $x_1, \dots, x_n$ from $\mathbb{P}(x|\theta)$ → **Algorithm** → $\hat{\theta}$

Parameter estimate

θ = unknown parameter

Maximum Likelihood Estimation (MLE). Given data $x_1, \dots, x_n$, find $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$ ("the MLE") such that $L(x_1, \dots, x_n | \hat{\theta})$ is maximized!

Likelihood – Continuous Case

Definition. The **likelihood** of independent observations $x_1, \dots, x_n$ is

$$L(x_1, \dots, x_n | \theta) = \prod_{i=1}^{n} f(x_i | \theta)$$

Example – Gaussian Parameters

Normal outcomes $x_1, \dots, x_n$, known variance $\sigma^2 = 1$

Goal: MLE for μ = expectation

$$L(x_1, \dots, x_n | \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2}} = \left(\frac{1}{\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2}}$$

$$\ln L(x_1, \dots, x_n | \mu) = -n \frac{\ln 2\pi}{2} - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2}$$

Goal: estimate μ = expectation

Example – Gaussian Parameters

$$\ln L(x_1, \dots, x_n|\mu) = -n\frac{\ln 2\pi}{2} - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2}$$

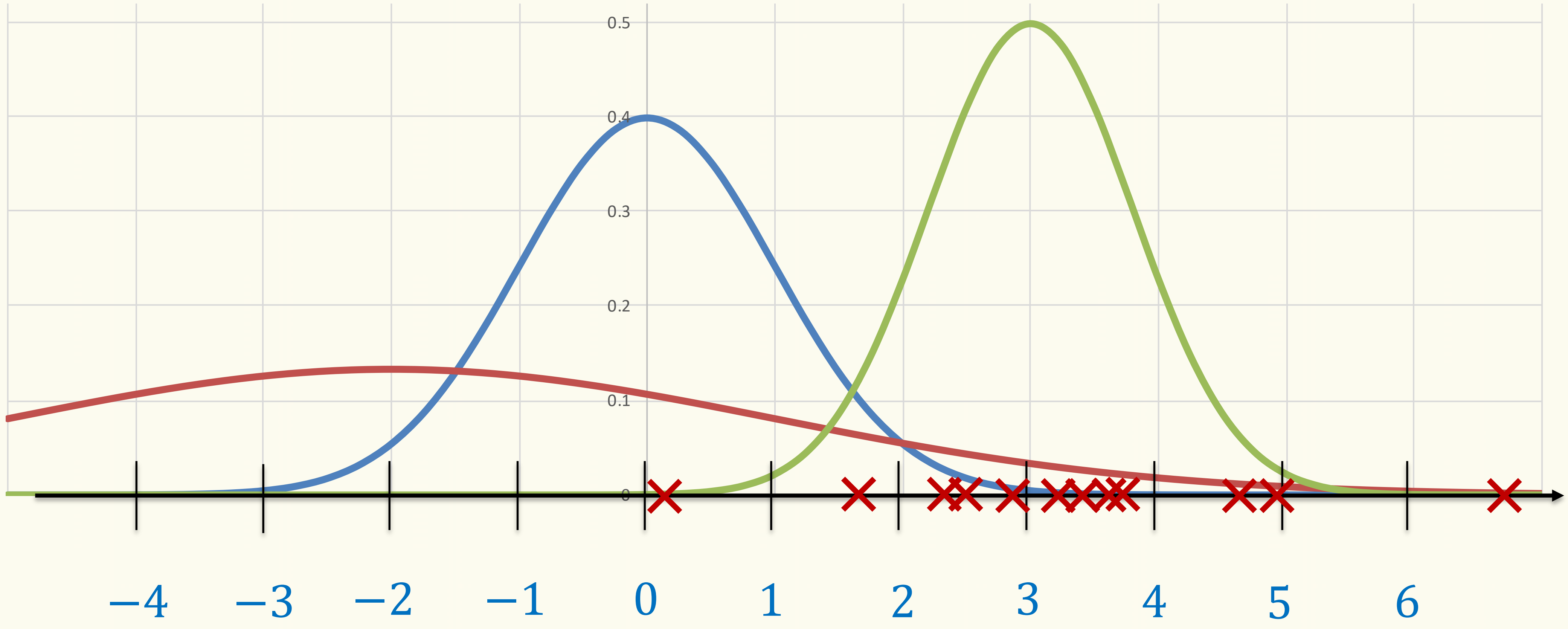
Note: $\frac{\partial}{\partial\mu}\frac{(x_i-\mu)^2}{2} = \frac{1}{2}\cdot 2\cdot(x_i-\mu)\cdot(-1) = \mu - x_i$

$$\frac{\partial}{\partial\mu}\ln L(x_1, \dots, x_n|\mu) = \sum_{i=1}^{n}(x_i-\mu) = \sum_{i=1}^{n} x_i - n\mu = 0$$

$$\hat{\mu} = \frac{\sum_i^n x_i}{n}$$

In other words, MLE is the *population mean* of the data.

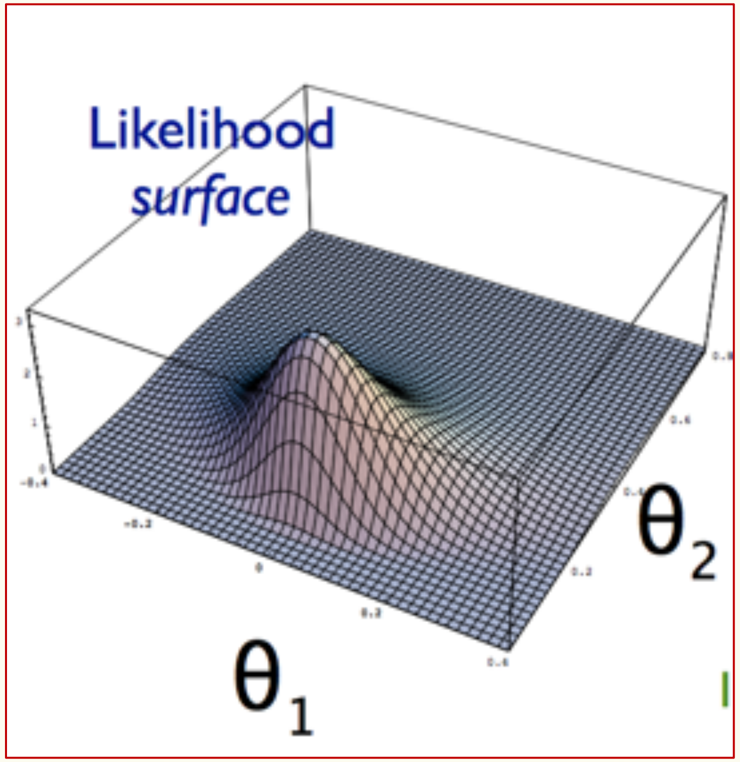
n samples $x_1, \dots, x_n \in \mathbb{R}$ from Gaussian $\mathcal{N}(\mu, \sigma^2)$. Most likely μ and σ^2?



Two-parameter optimization

Normal outcomes $x_1, \dots, x_n$

Goal: estimate $\theta_1 = \mu$ = expectation and $\theta_2 = \sigma^2$ = variance



$$L(x_1, \dots, x_n | \theta_1, \theta_2) = \left(\frac{1}{\sqrt{2\pi\theta_2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\theta_1)^2}{2\theta_2}}$$

$$\ln L(x_1, \dots, x_n | \theta_1, \theta_2) =$$

$$= -n\frac{\ln 2\pi\,\theta_2}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta_1)^2}{2\theta_2}$$

Two-parameter estimation

$$\ln L(x_1, \dots, x_n | \theta_1, \theta_2) = -\frac{\ln 2\pi \theta_2}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta_1)^2}{2\theta_2}$$

We need to find a solution $\hat{\theta}_1, \hat{\theta}_2$ to

$$\frac{\partial}{\partial \theta_1} \ln L(x_1, \dots, x_n | \theta_1, \theta_2) = 0$$
$$\frac{\partial}{\partial \theta_2} \ln L(x_1, \dots, x_n | \theta_1, \theta_2) = 0$$

MLE for Expectation

$$\ln L(x_1, \dots, x_n | \theta_1, \theta_2) = -n\frac{\ln 2\pi\,\theta_2}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta_1)^2}{2\theta_2}$$

$$\frac{\partial}{\partial \theta_1} \ln L(x_1, \dots, x_n | \theta_1, \theta_2) = \frac{1}{\theta_2}\sum_{i}^{n}(x_i - \theta_1) = 0$$

$$\boxed{\hat{\theta}_1 = \frac{\sum_i^n x_i}{n}}$$

In other words, MLE of expectation is (again) the *population mean* of the data, regardless of θ_2

What about the variance?

MLE for Variance

$$\ln L\left(x_1, \dots, x_n \middle| \hat{\theta}_1, \theta_2\right) = -n\frac{\ln 2\pi\, \theta_2}{2} - \sum_{i=1}^{n} \frac{\left(x_i - \hat{\theta}_1\right)^2}{2\theta_2}$$

$$= -n\frac{\ln 2\pi}{2} - n\frac{\ln \theta_2}{2} - \frac{1}{2\theta_2}\sum_{i=1}^{n}\left(x_i - \hat{\theta}_1\right)^2$$

$$\frac{\partial}{\partial \theta_2} \ln L\left(x_1, \dots, x_n \middle| \hat{\theta}_1, \theta_2\right) = -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_{i=1}^{n}\left(x_i - \hat{\theta}_1\right)^2 = 0$$

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\theta}_1\right)^2$$

In other words, MLE of variance is the *population variance* of the data.

So far

- We have decided that MLE estimators are always good.
- But why is it really the case?
  - Next: A natural property not always satisfied by MLE
  - And why MLE is nonetheless "good"

When is an estimator good?

Parameter estimate

Distribution $\mathbb{P}(x|\theta)$ → samples $X_1, \dots, X_n$ from $\mathbb{P}(x|\theta)$ → **Algorithm** → Θ_n

θ = unknown parameter

Definition. An estimator is **unbiased** if for all $n \geq 1$,

$$\mathbb{E}(\Theta_n) = \theta.$$

Example – Coin Flips

Recall: $\hat{\theta} = \frac{n_H}{n}$

Coin-flip outcomes $x_1, \ldots, x_n$, with n_H heads, n_T tails

Fact. $\hat{\theta}$ is unbiased

Let $Z_1, \ldots, Z_n$ be s.t. $Z_i = 1$ iff $x_i = H$ (and 0 otherwise)

In particular $\mathbb{P}(Z_i = 1) = \theta$

$$\widehat{\Theta} = \frac{1}{n}\sum_{i=1}^{n} Z_i \qquad \mathbb{E}(\widehat{\Theta}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(Z_i) = \frac{1}{n} n \cdot \theta = \theta$$

Notes

- Unbiasedness is not the ultimate goal either
  - Consider estimator which sets $\hat{\theta} = 1$ if first coin toss is heads, and $\hat{\theta} = 0$ otherwise – regardless of number of samples.
  - $\mathbb{P}(\Theta_n = 1) = \theta$
  - $\mathbb{E}(\Theta_n) = \theta$
- Generally, we would like instead $\mathbb{P}(\Theta_n \approx \theta)$ with high probability as $n \to \infty$.
  - Will discuss this on Monday.
  - Unbiasedness is a step towards this.

Example – Gaussian

Normal outcomes $X_1, \ldots, X_n$ iid according to $\mathcal{N}(\mu, \sigma^2)$

$$\widehat{\Theta}_1 = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\widehat{\Theta}_2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_1\right)^2$$

Example – Gaussian

Normal outcomes $X_1, \dots, X_n$ iid according to $\mathcal{N}(\mu, \sigma^2)$

$$\boxed{\widehat{\Theta}_1 = \frac{\sum_i^n X_i}{n}}$$

$$\mathbb{E}(\widehat{\Theta}_1) = \frac{\sum_{i=1}^n \mathbb{E}(X_i)}{n} = \frac{n \cdot \mu}{n} = \mu$$

Therefore: Unbiased!

Example – Gaussian

Normal outcomes $X_1, \ldots, X_n$ iid according to $\mathcal{N}(\mu, \sigma^2)$ Assume: $\sigma^2 > 0$

$$\widehat{\Theta}_2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_1\right)^2$$

$$\widehat{\Theta}_2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_1\right)^2$$

Unbiased!

Example: $n = 1$

$$\widehat{\Theta}_1 = \frac{X_1}{1} = X_1 \qquad \widehat{\Theta}_2 = \frac{1}{1}(X_1 - X_1)^2 = 0 \qquad \mathbb{E}(\widehat{\Theta}_2) = 0 \neq \sigma^2$$

Therefore: Biased!

Next time: Unbiased estimator proof + more intuition + confidence intervals

Example – Consistency

Normal outcomes $X_1, \dots, X_n$ iid according to $\mathcal{N}(\mu, \sigma^2)$ Assume: $\sigma^2 > 0$

$$\widehat{\Theta}_2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_1\right)^2$$

Population variance – <u>Biased!</u>

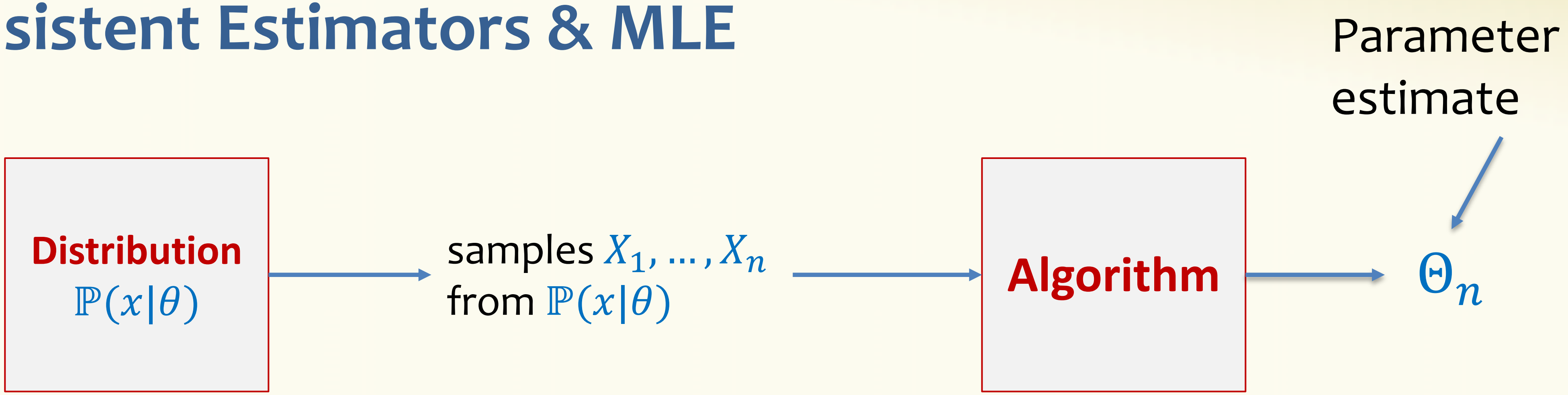
$$\widehat{\Theta}_2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_1\right)^2$$

Sample variance – <u>Unbiased!</u>

Left $\widehat{\Theta}_2$ converges to same value as right $\widehat{\Theta}_2$, i.e., σ^2, as $n \to \infty$.

Left $\widehat{\Theta}_2$ is "consistent"

Consistent Estimators & MLE



θ = unknown parameter

Definition. An estimator is **unbiased** if $\mathbb{E}(\Theta_n) = \theta$ for all $n \geq 1$.

Definition. An estimator is **consistent** if $\lim_{n \to \infty} \mathbb{E}(\Theta_\mathrm{n}) = \theta$.

Theorem. MLE estimators are consistent.

(But not necessarily unbiased)