Foundations of Inference: Risk and Loss Functions (Reading: Ch 10)

Learning goals:

- Introduce statistical risk as our method to evaluate an estimator.
- Understand why we use empirical risk instead of statistical risk.
- Perform estimation by minimizing a loss function.

UC Berkeley Data 100 Summer 2019
Sam Lau

(Slides adapted from Sandrine Dudoit and Joey Gonzalez, images adapted from David Quarfoot)

CC BY NC SA

Announcements

- Class is now harder if you don't have Stats background
  - 1st half of class: main practice comes in Lab
  - 2nd half of class: main practice comes in Discussion
- Mid-semester survey due **11:59pm tonight**
  - If ≥ 90% of class fills out, everyone gets 0.5 points added to midterm. Currently only 56 responses.
- Updated due dates:
  - HW4 due **Tuesday**
  - HW5 out next Tuesday, due next **Friday**

Statistical Risk

BY NC SA

Yesterday...

- Estimator is a numeric summary of a sample
- If sample was collected randomly, estimator is a RV
- To pick an estimator, we looked at its bias and variance
  - $B(\theta) = 0$ means estimator is accurate
  - $Var(\theta)$ low means estimator is precise

CC BY NC SA

An Analogy

- If archer hits the bullseye on one shot, is the archer good?
  - Maybe, but we're interested in repeated attempts

Archer → Estimator: $\hat{\theta}(X_1, \ldots, X_n)$

Bow and arrow → Sample data: $x_1, \ldots, x_n$

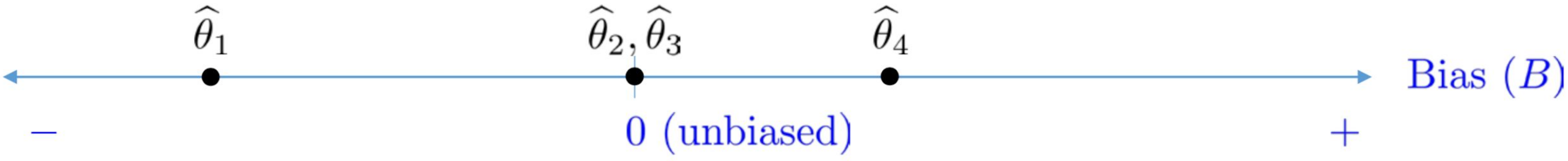
Shots in target → Estimate: $\hat{\theta}(x_1, \ldots, x_n)$

Bullseye → Parameter: θ^*

- Bias = 0 means archer's shots are centered at bullseye
- Variance low means archer's shots land close together

Picking an Estimator

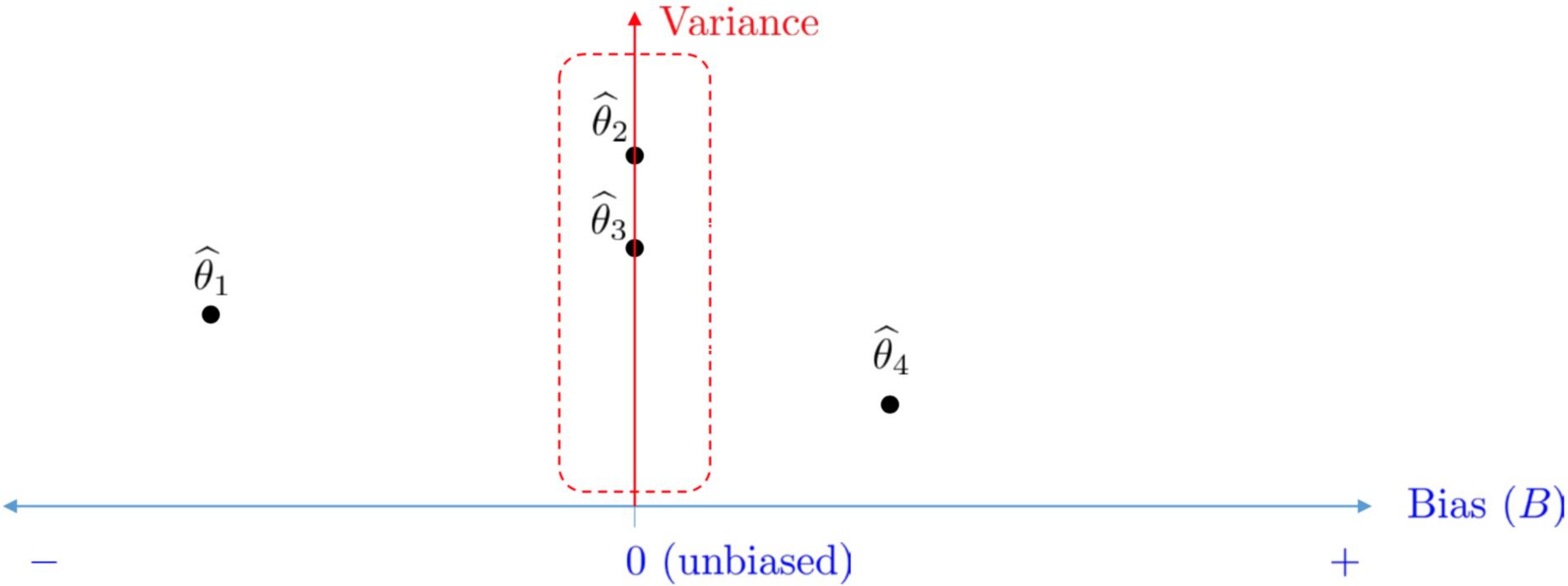
- Given a bunch of estimators, can find bias of each:



- What about estimators with zero bias?

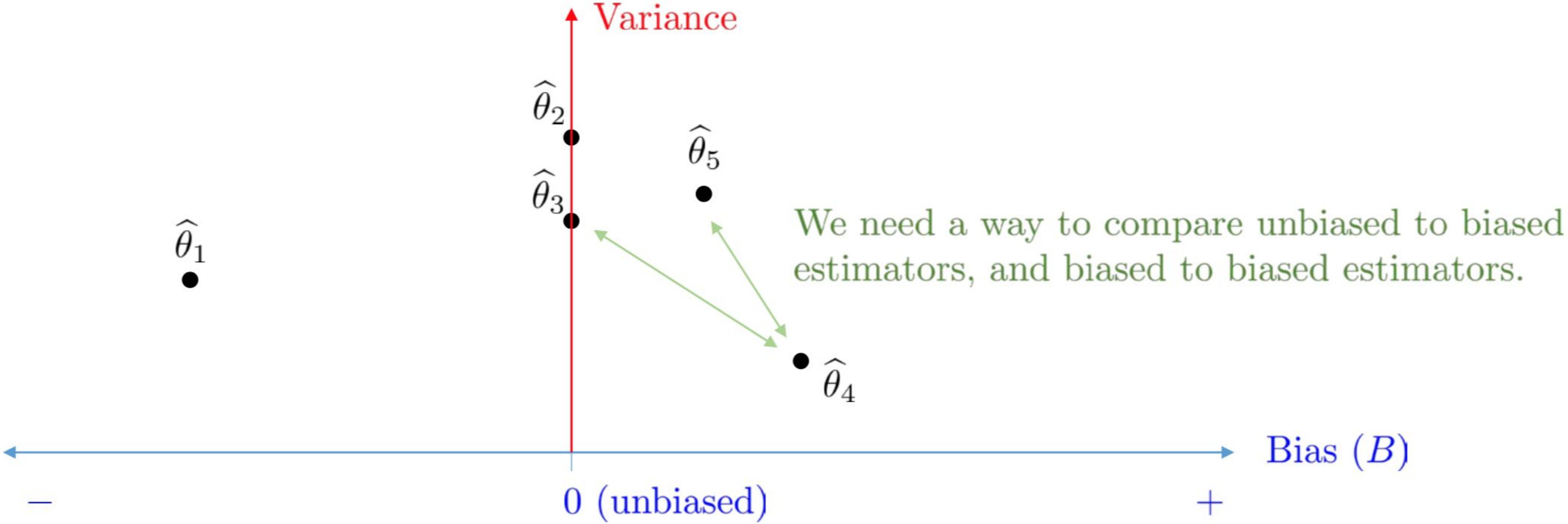
Picking an Estimator

- Many estimators have zero bias! Compare using variance:



Picking an Estimator

- What if θ is just a bit biased but has low variance?



Loss

- Let X be a random value drawn from the **population**.
- Let θ be an estimate for the population mean θ*.
  - θ is created from a single sample.
- The **squared error loss** of θ is:

$$L(\theta) = (X - \theta)^2$$

- Loss is a random variable with **higher** values when estimate wrong.

Risk

- **Statistical risk**: expected loss over all points in population.
- Let X be a random variable drawn from the **population**.
- Let θ be an estimate. The statistical risk of θ is:

$$R(\theta) = E\left[L(\theta)\right] = E\left[(X - \theta)^2\right]$$

- Think of the risk as an oracle: give it an estimate and the risk will tell you how good it is.
- Lower values = better!

CC BY NC SA

Risk combines bias and variance!

$$
\begin{aligned}
R(\theta) = E\left[L(\theta)\right] &= E\left[(X-\theta)^2\right] \\
&\dots && \text{(A bunch of algebra)} \\
&= (E[X]-\theta)^2 + E\left[(X-E[X])^2\right] \\
&= (\text{bias of } \theta)^2 + \text{Var}(X)
\end{aligned}
$$

- The risk gives a combined measure of bias and variance!
- For us, the statistically optimal estimator is the one with the **lowest risk**.
- This means that setting θ = E(X) is optimal!

CC BY NC SA

The Problem of E(X)

- Problem: We can't find E(X) easily. Why not?
- Instead, **estimate** E(X) using our sample.
  - Insight: a SRS looks like population, so we'll pretend that the sample **is** the population!

$$\widehat{E}(X) = \frac{1}{n}\sum x_i \approx E(X)$$

CC BY NC SA

Empirical Risk

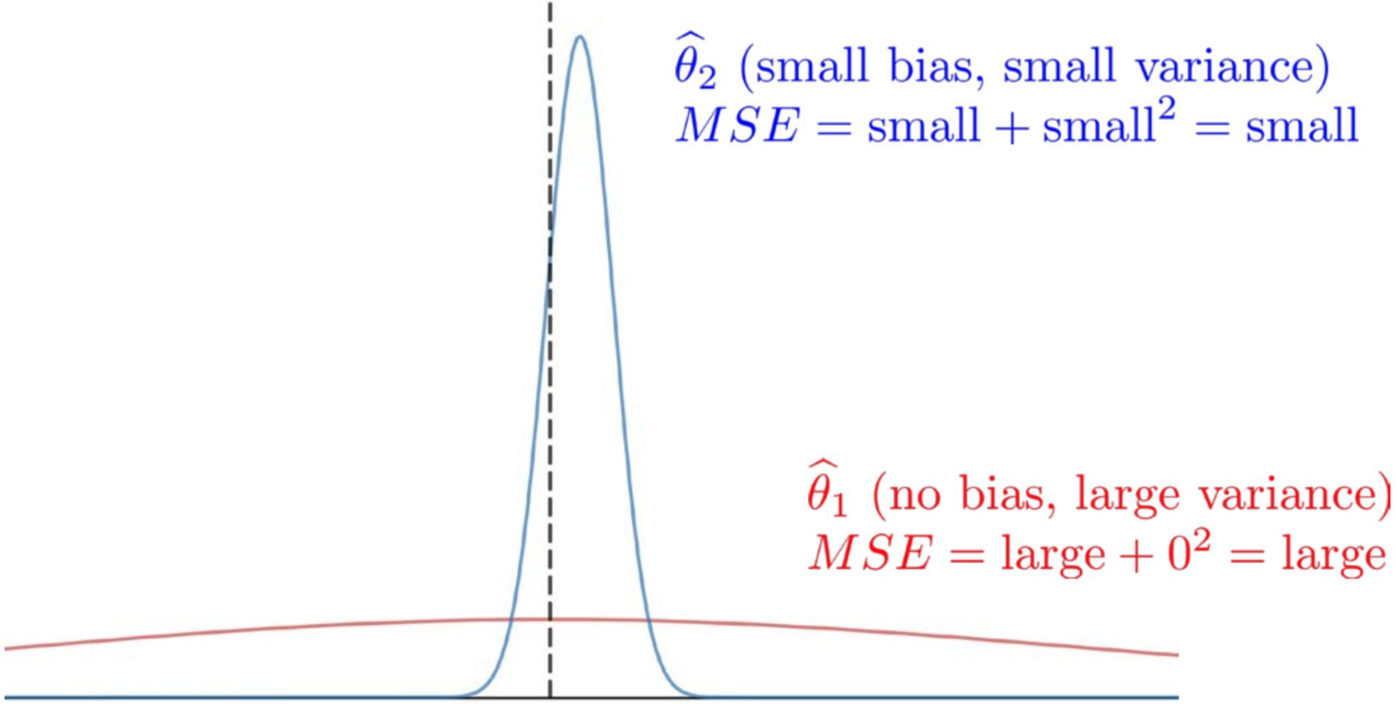
- Estimating E(X) in this way gives us the **empirical risk**.
- Calculated using our sample instead of the population:

$$\widehat{R}(\theta) = (\widehat{E}[X] - \theta)^2 + \widehat{Var}(X) \approx R(\theta)$$

- We can't minimize the risk, so we minimize the empirical risk instead.
- This means we should set θ = mean of sample! Why?

CC BY NC SA

Risk Visualized



Who is better?

A wildly inconsistent archer that hits the bullseye on average?

Or a very consistent archer that gets near the bullseye?

BY NC SA

Notation

- Let's go over notation:

θ^* : Population parameter (e.g. $\mu^*, (\sigma^2)^*$)

$\theta = \frac{1}{n}\sum X_i$: Estimat**or** (random variable, outputs estimates)

$\theta = \frac{1}{2}(x_1 + x_2)$: Estimate (not random, single value based on sample)

$\hat{\theta} = \frac{1}{n}\sum x_i$: Estimate that minimizes empirical risk

Great Expectations

- Tons of expectations! Here are the most important ones:

X is a random variable:
$E(X)$ is the long-run average of X

X is a RV drawn from the population:
$E(X) = \mu^*$ is the population mean.

θ is an estimator:
$E(\theta) - \theta^*$ is the bias of θ

$L(\theta)$ is a loss function, θ is an estimate:
$E\left[L(\theta)\right] = R(\theta)$ is the statistical risk.

CC BY NC SA

Break!

Fill out Attendance:

http://bit.ly/at-d100

cc BY NC SA

Loss Functions

Empirical Loss

- Recall: loss function measures an estimate θ against a random value drawn from population X.
- Mean squared error (MSE) loss:

$$L(\theta) = (X - \theta)^2$$

- **Average empirical loss** = substitute actual data for X and take the average:

$$L(\theta, x_1, \ldots, x_n) = \frac{1}{n} \sum (x_i - \theta)^2$$

- Notice that this is the same as the empirical risk!

Take the L

- Sample: $[1, 3, 5, 11]$
- Your estimate: $\theta = 4$
- What's the average MSE loss?

CC BY NC SA

Take the L

- Sample: $[1, 3, 5, 11]$
- Your estimate: $\theta = 4$
- What's the average MSE loss?

$$
\begin{aligned}
L(\theta, x_1, \ldots, x_n) &= \frac{1}{n} \sum (x_i - \theta)^2 \\
&= \frac{1}{4}(3^2 + 1^2 + 1^2 + 7^2) \\
&= 15
\end{aligned}
$$

CC BY NC SA

Minimizing the L

- We saw that average loss = empirical risk
  - It's more common to say "minimize loss"
- For the MSE loss, minimize loss with θ = sample mean
- We can also minimize loss with calculus:

$$L(\theta, x_1, \ldots, x_n) = \frac{1}{n} \sum (x_i - \theta)^2$$

$$\frac{\partial}{\partial \theta} L(\theta, x_1, \ldots, x_n) = \frac{1}{n} \sum (2)(x_i - \theta)(-1)$$

$$= -\frac{2}{n} \left(\sum (x_i) - n\theta \right)$$

$$\sum (x_i) - n\hat{\theta} = 0$$

$$\hat{\theta} = \frac{1}{n} \sum x_i$$

$$= x_i$$

CC BY NC SA

Your Turn

- Practice: prove that the minimizing θ for the MSE loss is the sample mean (don't peek at the previous slide!)

CC BY NC SA

Mean Absolute Error Loss

- Many other possible loss functions!
- **Mean absolute error (MAE)** loss function:

$$L(\theta, x_1, \ldots, x_n) = \frac{1}{n} \sum |x_i - \theta|$$

- Your turn: Prove that the minimizing θ = sample median.
- Assume that θ is not equal to any sample point.
- Hint: $$\frac{\partial}{\partial \theta} [|x|] = \text{sign}(x) \text{ , where } \text{sign}(x) = \begin{cases} 1 & \text{if } x > 0, \\ -1 & \text{if } x < 0 \end{cases}$$

Minimizing the MAE

$$L(\theta, \mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} |x_i - \theta|$$

$$= \frac{1}{n}\left(\sum_{x_i<\theta} |x_i - \theta| + \sum_{x_i=\theta} |x_i - \theta| + \sum_{x_i>\theta} |x_i - \theta|\right)$$

$$\frac{1}{n}\left(\sum_{x_i<\theta}(-1) + \sum_{x_i=\theta}(0) + \sum_{x_i>\theta}(1)\right) = 0$$

$$\sum_{x_i<\theta}(1) = \sum_{x_i>\theta}(1)$$

Wait, What?

$$\sum_{x_i<\theta}(1) = \sum_{x_i>\theta}(1)$$

- To minimize loss, pick θ so that the above equality holds.
- Notice that the LHS counts how many values are below θ, and the RHS counts how many values are above θ.
- This means we want θ to be the median of the sample!

CC BY NC SA

Putting it All Together

(demo)

CC BY NC SA

How do you pick a loss function?

- Depends on context and domain!
- MSE and MAE by far the most common
- MSE has severe penalties for very wrong values
  - Because of squared error
  - But sometimes this is what you want!
- E.g. pick MSE if you’re an airport and delays have cascading effects.

Summary

- Statistical risk gives us a metric to evaluate an estimator
  - Combines both bias and variance of estimator
- Since we can't compute statistical risk directly, use empirical risk / average empirical loss
  - Make estimations by minimizing the loss
- Two important loss functions today: MSE and MAE