Estimation:
Sample Averages, Bias, and Concentration Inequalities

CMPUT 267: Basics of Machine Learning

Logistics

Outline

1. Recap
2. Estimators
3. Concentration Inequalities
4. Consistency

Recap

- **Random variables** are functions from sample to some value
  - Upshot: A random variable takes different values with some probability
- The value of one variable can be informative about the value of another (because they are both functions of the same sample)
  - Distributions of multiple random variables are described by the **joint** probability distribution (joint PMF or joint PDF)
  - **Conditioning** on a random variable gives a new distribution over others
- X is **independent** of Y: conditioning on X does **not** give a new distribution over Y
  - X is **conditionally independent** of Y given Z:
    $P(Y \mid X, Z) = P(Y \mid Z); \qquad P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$

Recap

- **Bayes' Rule**

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)}$$

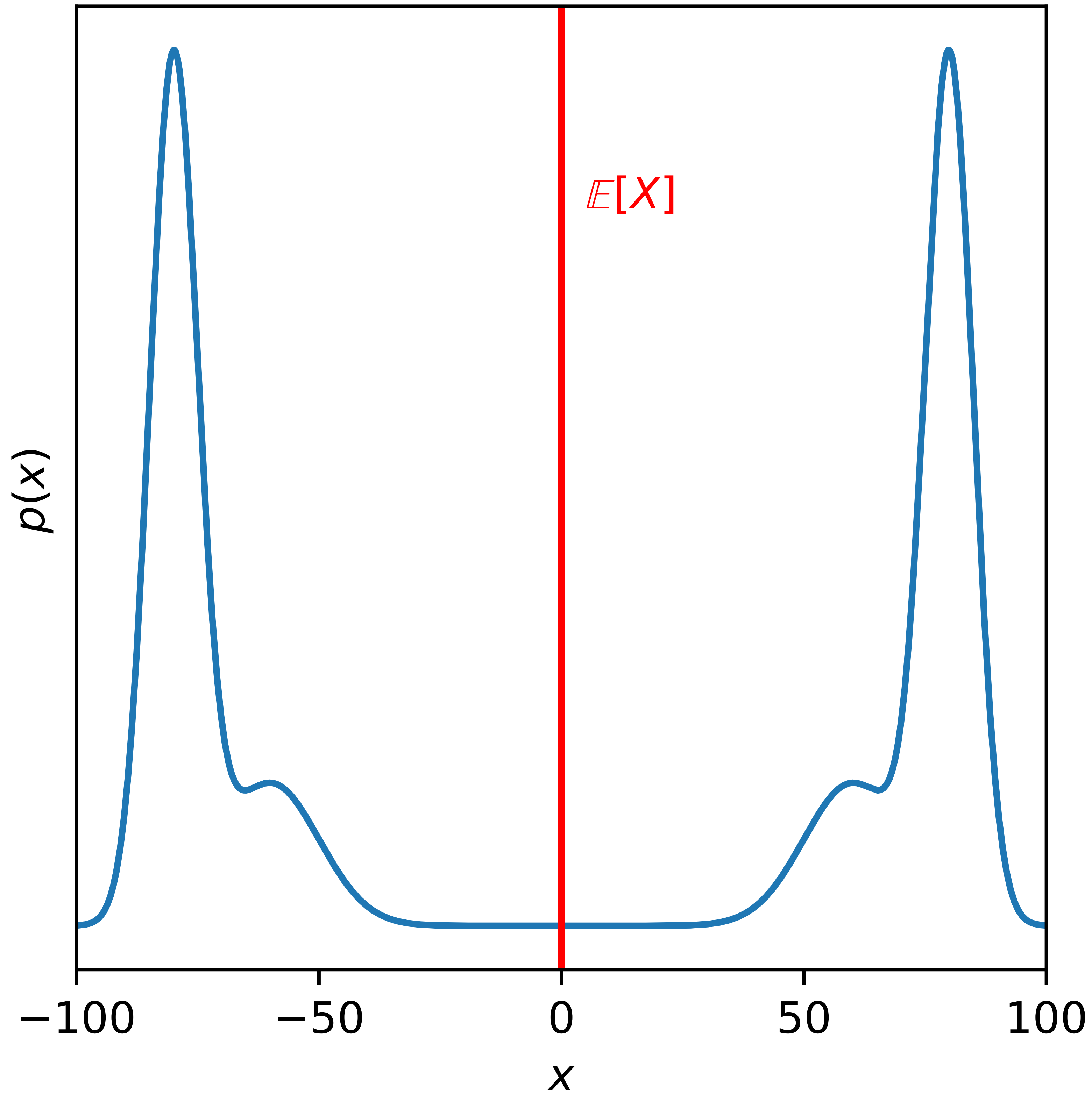
- The **expected value** of a random variable is an **average** over its values, **weighted** by the probability of each value

- The **variance** $\mathrm{Var}[X]$ of a random variable X is its expected squared distance from the mean

Estimators

Definition: An **estimator** is a procedure for estimating an unobserved quantity based on data.

Example: Estimating $\mathbb{E}[X]$ for r.v. $X \in \mathbb{R}$.



Questions:

random variable!

Suppose we can observe a different variable Y. Is Y a good estimator of $\mathbb{E}[X]$ in the following cases? Why or why not?

1. $Y \sim$ Uniform$[0,10]$
2. $Y = \mathbb{E}[X] + Z,$ where $Z \sim$ Uniform$[0,1]$
3. $Y = \mathbb{E}[X] + Z,$ where $Z \sim N(0,100^2)$
4. $Y = X$
5. How would you estimate $\mathbb{E}[X]$?

Estimators

- $\hat{X}$: the estimator
- How can we measure how good $\hat{X}$ is at estimating the true value?
- We can look at the properties of an estimator
- Expected value, variance
- A measure for how far $\hat{X}$ is from the true value.
  - The expected value of this measure
- **Bias**

Bias

Definition: The bias of an estimator $\hat{X}$ is its expected difference from the true value of the estimated quantity X:

$$\mathrm{Bias}(\hat{X}) = \mathbb{E}[\hat{X} - X]$$

- Bias can be positive or negative or zero
- When $\mathrm{Bias}(\hat{X}) = 0$, we say that the estimator $\hat{X}$ is **unbiased**

Bias

Definition: The bias of an estimator $\hat{X}$ is its expected difference from the true value of the estimated quantity X:

$$\text{Bias}(\hat{X}) = \mathbb{E}[\hat{X} - X]$$

Questions:

What is the bias of the following estimators of $\mathbb{E}[X]$?

1. $Y \sim$ Uniform$[0,10]$
2. $Y = \mathbb{E}[X] + Z$, where $Z \sim$ Uniform$[0,1]$
3. $Y = \mathbb{E}[X] + Z$, where $Z \sim N(0,100^2)$
4. $Y = X$

Independent and Identically Distributed (i.i.d.) Samples

- We usually won't try to estimate anything about a distribution based on only a single sample
- Usually, we use **multiple samples** from the **same distribution**
  - *Multiple samples:* This gives us more information
  - *Same distribution:* We want to learn about a single population
- One additional condition: the samples must be **independent** (**why?**)

Definition: When a set of random variables $X_1, X_2, \ldots$ are all independent, and each has the same distribution $X \sim F$, we say they are **i.i.d.** (independent and identically distributed), written

$$X_1, X_2, \ldots \overset{i.i.d.}{\sim} F.$$

Estimating Expected Value via the Sample Mean

Example: We have n i.i.d. samples from the same distribution F,

$$X_1, X_2, \ldots, X_n \overset{i.i.d}{\sim} F,$$

with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$ for each X_i.

We want to estimate μ.

Let's use the **sample mean** $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ to estimate μ.

Question: Is this estimator **unbiased**?

Question: Are **more samples** better? Why?

Estimating Expected Value via the Sample Mean

Example: We have n i.i.d. samples from the same distribution F,

$$X_1, X_2, \ldots, X_n \overset{i.i.d}{\sim} F,$$

with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$ for each X_i.

We want to estimate μ.

Let's use the **sample mean** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ to estimate μ.

Question: Is this estimator **unbiased**?

Question: Are **more samples** better? Why?

$$\begin{aligned}
\mathbb{E}[\bar{X}] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu \\
&= \frac{1}{n} n\mu \\
&= \mu \,. \quad \blacksquare
\end{aligned}$$

Estimating Expected Value via the Sample Mean

Example: Coin flip. X_i : value of coin flip i, $X_i \in \{0,1\}$, $X_i \sim$ Bernoulli, iid

$$\begin{aligned}\mathbb{E}[\bar{X}] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mu \\ &= \frac{1}{n}n\mu \\ &= \mu\,. \qquad \blacksquare\end{aligned}$$

Question: Is this estimator **unbiased**?

Question: Are **more samples** better? Why?

Variance of the Estimator

- Intuitively, more samples should make the estimator "closer" to the estimated quantity
- We can formalize this intuition partly by characterizing the **variance $\mathrm{Var}[\hat{X}]$ of the estimator itself**.
  - The variance of the estimator should decrease as the number of samples increases

Variance of the Estimator

- Intuitively, more samples should make the estimator "closer" to the estimated quantity
- We can formalize this intuition partly by characterizing the **variance $\mathrm{Var}[\hat{X}]$ of the estimator itself**.
  - The variance of the estimator should decrease as the number of samples increases
- **Example:** $\bar{X}$ for estimating μ:
  - The variance of the estimator shrinks linearly as the number of samples grows.

$$
\begin{aligned}
\mathrm{Var}[\bar{X}] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[X_i] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 \\
&= \frac{1}{n^2}n\sigma^2 \quad = \frac{1}{n}\sigma^2.
\end{aligned}
$$

Variance of the Estimator

- Intuitively, more samples should make the estimator "closer" to the estimated quantity
- We can formalize this intuition partly by characterizing the **variance $\text{Var}[\hat{X}]$ of the estimator itself**.
  - The variance of the estimator should decrease as the number of samples increases
- **Example:** $\bar{X}$ for estimating μ:
  - The variance of the estimator shrinks linearly as the number of samples grows.

$$
\begin{aligned}
\text{Var}[\bar{X}] &= \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n^2}\text{Var}\left[\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}[X_i] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 \\
&= \frac{1}{n^2}n\sigma^2 \quad = \frac{1}{n}\sigma^2.
\end{aligned}
$$

Variance of the Estimator

- **Example:** $\bar{X}$ for estimating μ:
  - The variance of the estimator shrinks **linearly** as the number of samples grows.
  - $\hat{X} \xrightarrow{n \to \infty} \mu$
- For finite n, how good of an estimate is $\hat{X}$?
- We want the difference from the expected value to be **small**, and be **consistently** small - we want to obtain a confidence interval around our estimate.

$$
\begin{aligned}
\mathrm{Var}[\bar{X}] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[X_i] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 \\
&= \frac{1}{n^2}n\sigma^2 \quad = \frac{1}{n}\sigma^2.
\end{aligned}
$$

Confidence intervals

- We want to obtain a confidence interval around our estimate - we want the difference from the expected value to be small, and be consistently small.
- We would like to be able to claim $\Pr\left(\left|\bar{X}-\mu\right| < \epsilon\right) > 1-\delta$ for some $\delta, \epsilon > 0$
- This tells us that $\mathbb{E}[\bar{X}] \in \{\bar{X}-\epsilon, \bar{X}+\epsilon\}$ with a large probability, $1-\delta$
- Confidence level: δ, width of interval: ϵ
- $\Pr\left(\left|\bar{X}-\mu\right| < \epsilon\right) > 1-\delta$ for *any* $\delta, \epsilon > 0$ that we pick (**why?**)
- $\mathrm{Var}[\bar{X}] = \dfrac{1}{n}\sigma^2$ means that with "enough" data we can get close to the expected value.
- Suppose we have $n = 10$ samples, and we know $\sigma^2 = 81$; so $\mathrm{Var}[\bar{X}] = 8.1$.
- **Question:** What is $\Pr\left(\left|\bar{X}-\mu\right| < 2\right)$?

Variance Is Not Enough

Knowing $\mathrm{Var}[\bar{X}] = 8.1$ is **not enough** to compute $\Pr(|\bar{X} - \mu| < 2)$!

Examples:

$$p(\bar{x}) = \begin{cases} 0.9 & \text{if } \bar{x} = \mu \\ 0.05 & \text{if } \bar{x} = \mu \pm 9 \end{cases} \implies \mathrm{Var}[\bar{X}] = 8.1 \text{ and } \Pr(|\bar{X} - \mu| < 2) = 0.9$$

$$p(\bar{x}) = \begin{cases} 0.999 & \text{if } \bar{x} = \mu \\ 0.0005 & \text{if } \bar{x} = \mu \pm 90 \end{cases} \implies \mathrm{Var}[\bar{X}] = 8.1 \text{ and } \Pr(|\bar{X} - \mu| < 2) = 0.999$$

$$p(\bar{x}) = \begin{cases} 0.1 & \text{if } \bar{x} = \mu \\ 0.45 & \text{if } \bar{x} = \mu \pm 3 \end{cases} \implies \mathrm{Var}[\bar{X}] = 8.1 \text{ and } \Pr(|\bar{X} - \mu| < 2) = 0.1$$

Hoeffding's Inequality

Theorem: Hoeffding's Inequality

Suppose that $X_1, \ldots, X_n$ are distributed i.i.d, with $a \leq X_i \leq b$.
Then for any $\epsilon > 0$,

$$\Pr\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right).$$

Equivalently, $\Pr\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \leq (b-a)\sqrt{\frac{\ln(2/\delta)}{2n}}\right) \geq 1 - \delta.$

Chebyshev's Inequality

Theorem: Chebyshev's Inequality

Suppose that $X_1, \ldots, X_n$ are distributed i.i.d. with variance σ^2.
Then for any $\epsilon > 0$,

$$\Pr\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Equivalently, $\Pr\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \leq \sqrt{\frac{\sigma^2}{\delta n}}\right) \geq 1 - \delta.$

When to Use Chebyshev, When to Use Hoeffding?

- **Popoviciu's inequality:** If $a \le X_i \le b$, then $\mathrm{Var}[X_i] \le \frac{1}{4}(b-a)^2$
- Hoeffding's inequality gives $\epsilon = (b-a)\sqrt{\frac{\ln(2/\delta)}{2n}} = \sqrt{\frac{\ln(2/\delta)}{2}}(b-a)\sqrt{\frac{1}{n}}$;

  Chebyshev's inequality gives $\epsilon = \sqrt{\frac{\sigma^2}{\delta n}} \le \sqrt{\frac{(b-a)^2}{4\delta n}} = \frac{1}{2\sqrt{\delta}}(b-a)\sqrt{\frac{1}{n}}$
- **Hoeffding's inequality** gives a **tighter bound***, but it can only be used on **bounded** random variables
  - * whenever $\sqrt{\frac{\ln(2/\delta)}{2}} < \frac{1}{2\sqrt{\delta}} \iff \delta < \sim 0.232$
- **Chebyshev's inequality** can be applied even for **unbounded** variables

Consistency

Definition: A sequence of random variables X_n **converges in probability** to a random variable X (written $X_n \xrightarrow{p} X$) if for all $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr(|X_n - X| > \epsilon) = 0.$$

Definition: An estimator $\hat{X}$ for a quantity X is **consistent** if $\hat{X} \xrightarrow{p} X$.

Weak Law of Large Numbers

Theorem: Weak Law of Large Numbers

Let $X_1, \dots, X_n$ be distributed i.i.d. with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$.

Then the **sample mean**

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is a **consistent estimator** for μ.

Proof:

1. We have already shown that $\mathbb{E}[\bar{X}] = \mu$
2. By Chebyshev,
   $$\Pr\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}$$
   for arbitrary $\epsilon > 0$
3. Hence $\lim_{n \to \infty} \Pr\left(\left|\bar{X} - \mu\right| \geq \epsilon\right) = 0$
   for any $\epsilon > 0$
4. Hence $\bar{X} \xrightarrow{p} \mu$. ∎

Convergence Rate via Chebyshev

The **convergence rate** indicates how quickly the error in an estimator decays as the number of samples grows.

Example: Estimating mean of a distribution using $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

- Recall that **Chebyshev's inequality** guarantees

$$\Pr\left(\left| \bar{X} - \mathbb{E}[\bar{X}] \right| \leq \sqrt{\frac{\sigma^2}{\delta n}} \right) \geq 1 - \delta$$

- Convergence rate is thus $O\left(1/\sqrt{n}\right)$

Sample Complexity

Definition:
The **sample complexity** of an estimator is the number of samples required to guarantee an expected error of at most ϵ with probability $1 - \delta$, for given δ and ϵ.

- We want sample complexity to be small (**why?**)
- Sample complexity is determined by:
  1. The **estimator** itself
     - Smarter estimators can sometimes improve sample complexity
  2. Properties of the **data generating process**
     - If the data are high-variance, we need more samples for an accurate estimate
     - But we can reduce the sample complexity if we can **bias** our estimate **toward the correct value**

Sample Complexity

Definition:
The **sample complexity** of an estimator is the number of samples required to guarantee an expected error of at most ϵ with probability $1 - \delta$, for given δ and ϵ.

For $\delta = 0.05$, **Chebyshev** gives

$$\epsilon = \sqrt{\frac{\sigma^2}{\delta n}} = \frac{1}{\sqrt{0.05}} \frac{\sigma}{\sqrt{n}}$$

$$\iff \epsilon = 4.47 \frac{\sigma}{\sqrt{n}}$$

$$\iff \sqrt{n} = 4.47 \frac{\sigma}{\epsilon}$$

$$\iff n = 19.98 \frac{\sigma^2}{\epsilon^2}$$

With **Gaussian assumption** and $\delta = 0.05$,

$$\epsilon = 1.96 \frac{\sigma}{\sqrt{n}}$$

$$\iff \sqrt{n} = 1.96 \frac{\sigma}{\epsilon}$$

$$\iff n = 3.84 \frac{\sigma^2}{\epsilon^2}$$

How good is an estimator?

- **Bias:** whether an estimator is correct **in expectation**
- **Consistency:** whether an estimator is correct **in the limit of infinite data**
- **Convergence rate:** how fast the estimator **approaches its own mean**
  - For an **unbiased** estimator, this is also how fast its **error bounds** shrink
- We don't necessarily care about an estimator's being unbiased.
  - Often, what we care about is our estimator's **accuracy in expectation**

Mean-Squared Error

- We don't necessarily care about an estimator's being unbiased.
  - Often, what we care about is our estimator's **accuracy in expectation**

Definition: Mean squared error of an estimator $\hat{X}$ of a quantity X:

$$\text{MSE}(\hat{X}) = \mathbb{E}\left[(\hat{X} - \mathbb{E}[X])^2\right]$$

different!

Bias-Variance Decomposition

Sometimes a biased estimator can be closer to the estimated quantity than an unbiased one.

$$
\begin{aligned}
MSE(\hat{X}) &= \mathbb{E}[(\hat{X} - \mathbb{E}[X])^2] = \mathbb{E}[(\hat{X} - \mu)^2] && \mu = \mathbb{E}[X] \\
&= \mathbb{E}[(\hat{X} - \mathbb{E}[\hat{X}] + \mathbb{E}[\hat{X}] - \mu)^2] && -\mathbb{E}[\hat{X}] + \mathbb{E}[\hat{X}] = 0 \\
&= \mathbb{E}[((\hat{X} - \mathbb{E}[\hat{X}]) + b)^2] && b = \mathrm{Bias}(\hat{X}) = \mathbb{E}[\hat{X}] - \mu \\
&= \mathbb{E}[(\hat{X} - \mathbb{E}[\hat{X}])^2 + 2b(\hat{X} - \mathbb{E}[\hat{X}]) + b^2] \\
&= \mathbb{E}[(\hat{X} - \mathbb{E}[\hat{X}])^2] + \mathbb{E}[2b(\hat{X} - \mathbb{E}[\hat{X}])] + \mathbb{E}[b^2] && \text{linearity of } \mathbb{E} \\
&= \mathbb{E}[(\hat{X} - \mathbb{E}[\hat{X}])^2] + 2b\mathbb{E}[(\hat{X} - \mathbb{E}[\hat{X}])] + b^2 && \text{constants come out of } \mathbb{E} \\
&= \mathrm{Var}[\hat{X}] + 2b\mathbb{E}[(\hat{X} - \mathbb{E}[\hat{X}])] + b^2 && \text{def. variance} \\
&= \mathrm{Var}[\hat{X}] + 2b(\mathbb{E}[\hat{X}] - \mathbb{E}[\hat{X}]) + b^2 && \text{linearity of } \mathbb{E} \\
&= \mathrm{Var}[\hat{X}] + b^2 \\
&= \mathrm{Var}[\hat{X}] + \mathrm{Bias}(\hat{X})^2 \qquad \blacksquare
\end{aligned}
$$

Bias-Variance Tradeoff

$$\boxed{\mathrm{MSE}(\hat{X}) = \mathrm{Var}[\hat{X}] + \mathrm{Bias}(\hat{X})^2}$$

- If we can decrease bias without increasing variance, error goes down
- If we can decrease variance without increasing bias, error goes down
- **Question:** Would we ever want to **increase bias**?
- *YES.* If we can increase (squared) bias in a way that **decreases variance more**, then error goes down!
  - **Interpretation:** Biasing the estimator toward values that are **more likely to be true** (based on **prior information)**

Downward-biased Mean Estimation

Example: Let's estimate μ given i.i.d $X_1, \ldots, X_n$ with $\mathbb{E}[X_i] = \mu$ using: $Y = \dfrac{1}{n+100} \sum_{i=1}^{n} X_i$

This estimator is **biased**:

$$\mathbb{E}[Y] = \mathbb{E}\left[\frac{1}{n+100}\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{n+100}\sum_{i=1}^{n} \mathbb{E}[X_i]$$

$$= \frac{n}{n+100}\mu$$

$$\mathrm{Bias}(Y) = \frac{n}{n+100}\mu - \mu = \frac{-100}{n+100}\mu$$

This estimator has **low variance**:

$$\mathrm{Var}(Y) = \mathrm{Var}\left[\frac{1}{n+100}\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{(n+100)^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{(n+100)^2}\sum_{i=1}^{n} \mathrm{Var}[X_i]$$

$$= \frac{n}{(n+100)^2}\sigma^2$$

Estimating μ Near 0

Example: Suppose that $\sigma = 1$, $n = 10$, and $\mu = 0.1$

$\text{Bias}(\bar{X}) = 0$

$$\text{MSE}(\bar{X}) = \text{Var}(\bar{X}) + \text{Bias}(\bar{X})^2$$
$$= \text{Var}(\bar{X})$$

$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$

$$= \frac{1}{10}$$

$$\text{MSE}(Y) = \text{Var}(Y) + \text{Bias}(Y)^2$$
$$= \frac{n}{(n+100)^2}\sigma^2 + \left(\frac{100}{n+100}\mu\right)^2$$
$$= \frac{10}{110^2} + \left(\frac{100}{110}0.1\right)^2$$
$$\approx 9 \times 10^{-4}$$

Summary

- The **variance** $\mathrm{Var}[X]$ of a random variable X is its expected squared distance from the mean
- An **estimator** is a random variable representing a procedure for estimating the value of an unobserved quantity based on observed data
- **Concentration inequalities** let us bound the probability of a given estimator being at least ϵ from the estimated quantity
- An estimator is **consistent** if it **converges in probability** to the estimated quantity