Estimation: Sample Complexity and the Bias-Variance Tradeoff

CMPUT 296: Basics of Machine Learning

Textbook §3.4-3.5

Logistics

Reminders:

- Thought Question 1 (due **Thursday, September 17**)
- Assignment 1 (due **Thursday, September 24**)

Recap

- 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

When to Use Chebyshev, When to Use Hoeffding?

Popoviciu's inequality: If $a \leq X_i \leq b$, then $\text{Var}[X_i] \leq \frac{1}{4}(b-a)^2$

Hoeffding's inequality: $\epsilon = (b-a)\sqrt{\frac{\ln(2/\delta)}{2n}} \qquad = \sqrt{\frac{\ln(2/\delta)}{2}}(b-a)\sqrt{\frac{1}{n}}$

Chebyshev's inequality: $\epsilon = \sqrt{\frac{\sigma^2}{\delta n}} \leq \sqrt{\frac{(b-a)^2}{4\delta n}} \qquad = \frac{1}{2\sqrt{\delta}}(b-a)\sqrt{\frac{1}{n}}$

- Hoeffding's inequality gives a **tighter bound***, but it can only be used on **bounded** random variables
  - * whenever $\sqrt{\frac{\ln(2/\delta)}{2}} < \frac{1}{2\sqrt{\delta}}$
  - * E.g., if $\text{Var}[X_i] \approx \frac{1}{4}(b-a)^2$, then whenever $\delta < \sim 0.232$
- Chebyshev's inequality can be applied even for **unbounded** variables
  - or for bounded variables with known, small σ^2

Outline

1. Recap & Logistics
2. Sample Complexity
3. Bias-Variance Tradeoff

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**

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 $O\left(1/\sqrt{n}\right)$ (**why?**)

Convergence Rate via Gaussian

Example: Now assume that we know $X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, and we know σ^2 but not μ.

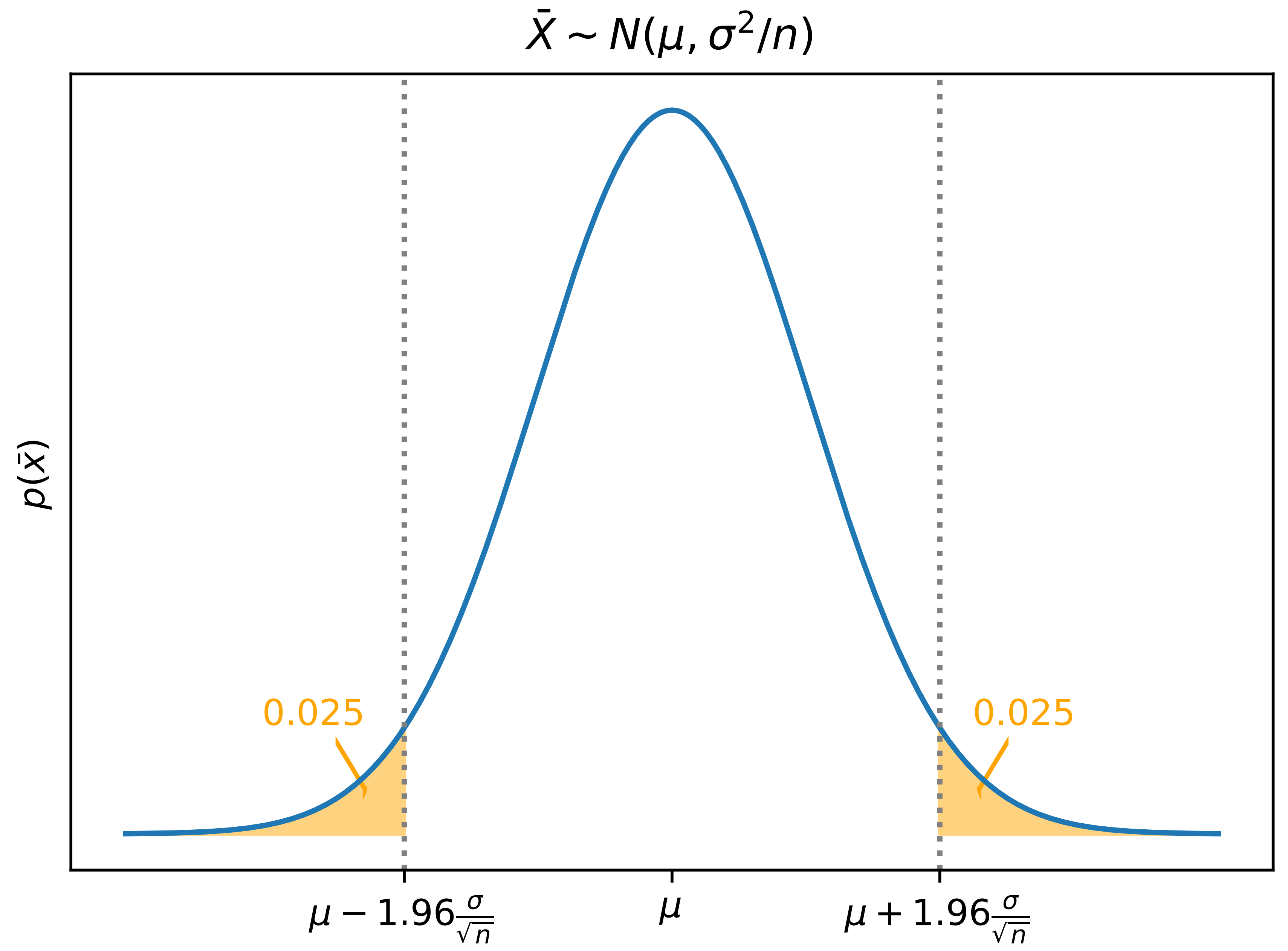
$$\bar{X} \sim N(\mu, \sigma^2/n)$$

Find ϵ such that $\Pr(|\bar{X} - \mu| < \epsilon) = 0.95$ by finding ϵ such that $\int_{-\infty}^{\epsilon} p(x)\,dx = 0.025$ (**why?**)

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

inverse CDF

```
__main__> import scipy.stats
__main__> scipy.stats.norm.ppf(0.025)
-1.959963984540054
__main__> scipy.stats.norm.cdf(-1.96)
0.024997895148220435
```



Questions:

1. What is the expected value of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$?
2. What is the variance of $\bar{X}$?
3. What is the distribution of $\bar{X}$?

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

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

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

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

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

With Gaussian assumption and $\delta = 0.05$,

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

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

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

Mean-Squared Error

- **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**

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

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

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

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

- 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: $Y = \frac{1}{n+100} \sum_{i=1}^{n} X_i$

This estimator is **biased**:

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

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

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

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

This estimator has **low variance**:

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

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

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

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

Estimating μ Near 0

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

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

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

$$= \mathrm{Var}(\bar{X})$$

$\mathrm{Var}(\bar{X}) = \dfrac{\sigma^2}{n}$

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

$$\mathrm{MSE}(Y) = \mathrm{Var}(Y) + \mathrm{Bias}(Y)^2$$

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

$$= \frac{10}{110^2} + \left(\frac{100}{110}0.1\right)^2$$

$$\approx 9 \times 10^{-4}$$

Prior Information and Bias: There's No Free Lunch

Example: Suppose that $\sigma = 1$, $n = 10$, and $\mu = 5$

$$\begin{aligned}
\text{MSE}(\bar{X}) &= \text{Var}(\bar{X}) + \text{Bias}(\bar{X})^2 \\
&= \text{Var}(\bar{X}) \\
&= \frac{1}{10}
\end{aligned}$$

$$\begin{aligned}
\text{MSE}(Y) &= \text{Var}(Y) + \text{Bias}(Y)^2 \\
&= \frac{n}{(n+100)^2}\sigma^2 + \left(\frac{-100}{n+100}\mu\right)^2 \\
&= \frac{10}{110^2} + \left(-\frac{100}{110}5\right)^2 \\
&\approx 20.66
\end{aligned}$$

Whoa! What went wrong?

Summary

- **Sample complexity** is the **number of samples** needed to attain a desired error bound ϵ at a desired probability $1 - \delta$
- The **mean squared error** of an estimator **decomposes** into **bias** (squared) and **variance**
- Using a **biased** estimator can have **lower error** than an unbiased estimator
    - Bias the estimator based some **prior information**
    - *But this only helps if the prior information is* ***correct***
    - Cannot reduce error by adding in arbitrary bias