Review for Midterm Exam 1

Chapter 2 (Probability)
Chapter 3 (Estimation):
Bias, Variance, Concentration Inequalities
Chapter 4 Optimization

CMPUT 267: Basics of Machine Learning

Logistics

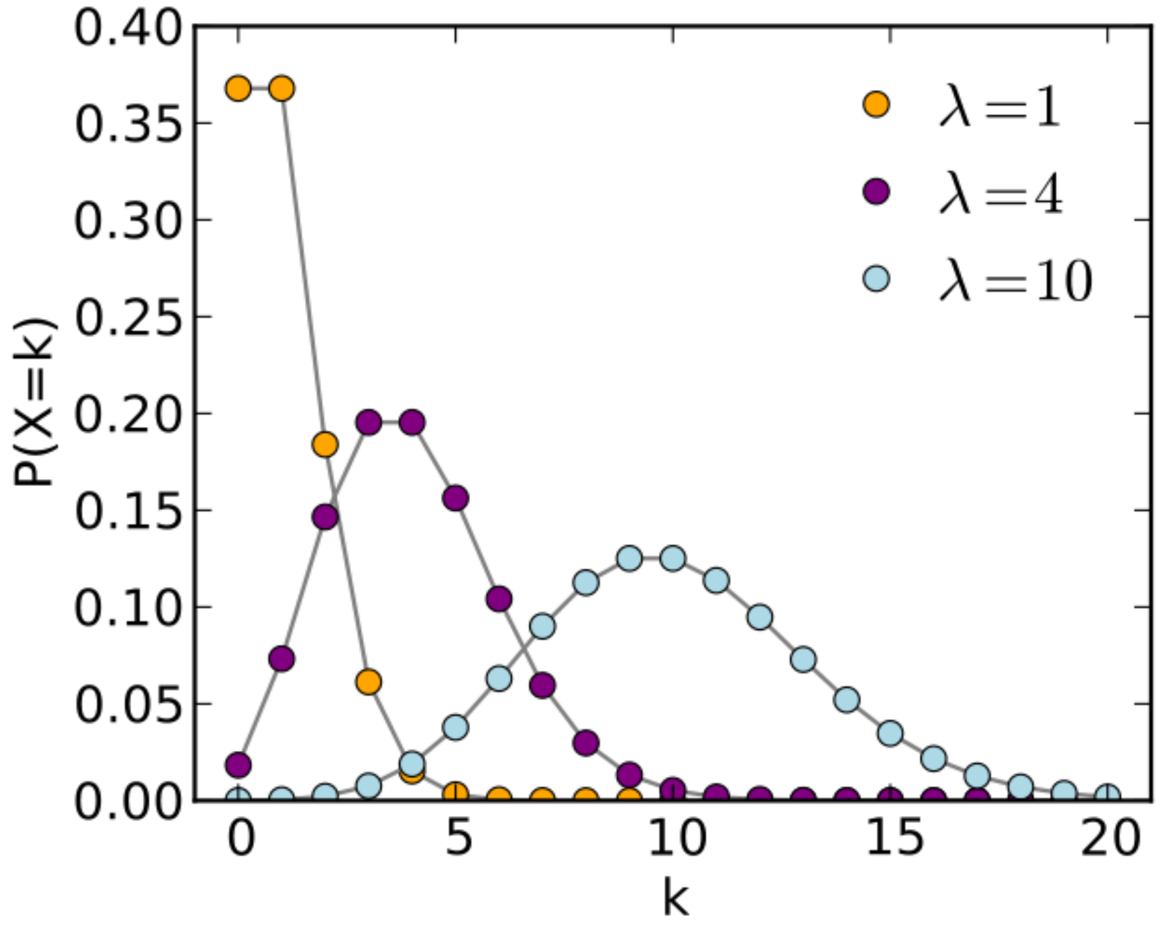
- Midterm Exam 1 **during class on Thursday Feb 15 in the usual classroom**
- Formula sheet provided (on the course site already). It will be printed for you - do not bring your own.
- The practice exam and the real exam are similar. Please review the practice exam!
    - But they are definitely not the same. Do not simply try to pattern match. You need to understand the practice exam, and be able to apply that knowledge.
    - The exam is meant to test the basics, not to challenge you; answers can be short (the exam is short so each question is worth a lot)

Language of Probabilities

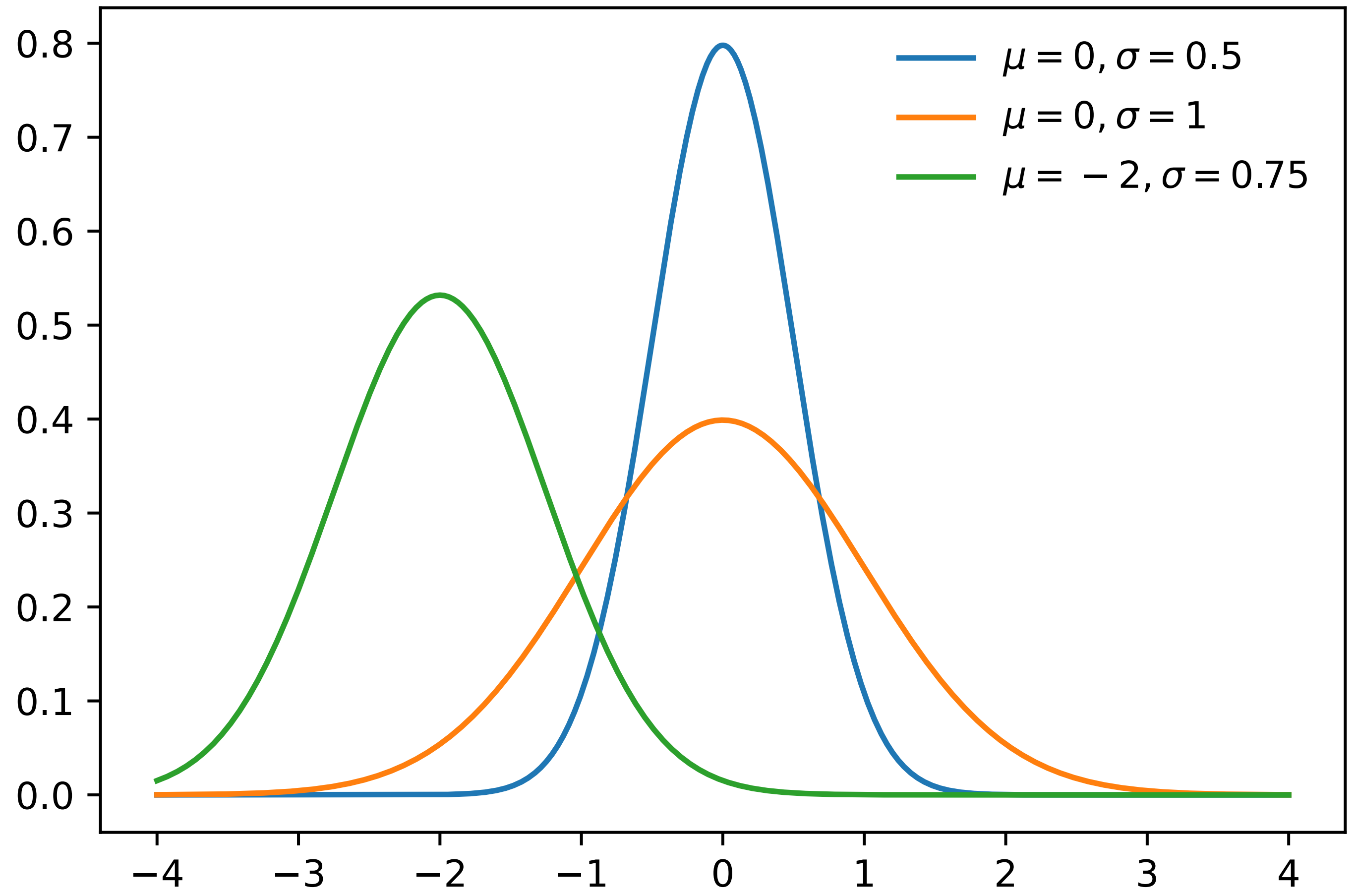
- Define random variables, and their distributions
  - So that we can formally reason about data and estimators
- Express our beliefs about behaviour of these RVs, and relationships to other RVs
- Examples:
  - p(x) Gaussian means we believe X is Gaussian distributed
  - p(y | X = x)—or written p(y | x)— is Gaussian means that when conditioned on x, y is Gaussian
  - p(w) and p(w | Data)

PMFs and PDFs

- Discrete RVs have PMFs
  - outcome space: e.g, $\Omega = \{1,2,3,4,5,6\}$
  - examples pmfs: probability tables, Poisson $p(k) = \dfrac{\lambda^k e^{-\lambda}}{k!}$



- Continuous RVs have PDFs
  - outcome space: e.g., $\Omega = [0,1]$
  - example pdf: Gaussian, Gamma



A few questions

- Do PMFs p(x) have to output values between [0,1]? Yes
- Do PDFs p(x) have to output values between [0,1]? No (between [0, infinity))
- What other condition(s) are put on a function p to make it a valid pmf or pdf?

A few questions

- Do PMFs p(x) have to output values between [0,1]? Yes
- Do PDFs p(x) have to output values between [0,1]? No (between [0, infinity))
- What other condition(s) are put on a function p to make it a valid pmf or pdf?
- PMF: $\sum_{x \in \mathcal{X}} p(x) = 1$
- PDF: $\int_{\mathcal{X}} p(x) dx = 1$

A few questions

- Is the following function a pdf or a pmf?
- $p(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b, \\ 0 & \text{otherwise.} \end{cases}$ i.e., $p(x) = \dfrac{1}{b-a}$ for $x \in [a, b]$

How would you define a uniform distribution for a discrete RV

How would you define a uniform distribution for a discrete RV

- Imagine $x \in \{1,2,3,4,5\}$
- What is the uniform pmf for this outcome space?

How would you define a uniform distribution for a discrete RV

- Imagine $x \in \{1,2,3,4,5\}$
- What is the uniform pmf for this outcome space?
- $$p(x) = \begin{cases} \frac{1}{5} & \text{if } x \in \{1,2,3,4,5\}, \\ 0 & \text{otherwise.} \end{cases}$$

How do you answer this probabilistic question?

How do you answer this probabilistic question?

- For continuous RV X with a uniform distribution and outcome space [0,10], what is the probability that X is greater than 7?

How do you answer this probabilistic question?

- For continuous RV X with a uniform distribution and outcome space [0,10], what is the probability that X is greater than 7?

$$\text{Pr}(X > 7) = \int_7^{10} p(x)dx = \int_7^{10} \frac{1}{10}dx$$

$$= \frac{1}{10}\int_7^{10} dx = \frac{1}{10}x\,|_7^{10}$$

$$= \frac{3}{10}$$

-

Multivariate Setting

- Conditional distribution, $p(y \mid x) = \dfrac{p(x, y)}{p(x)}$, Marginal $p(y) = \sum_{x \in \mathcal{X}} p(x, y)$
- Chain Rule $p(x, y) = p(y \mid x)p(x) = p(x \mid y)p(y)$
- Bayes Rule $p(y \mid x) = \dfrac{p(x \mid y)p(y)}{p(x)}$
- Law of total probability $p(y) = \sum_{x \in \mathcal{X}} p(y|x)p(x)$
- **Question**: How do you get the law of total probability from the chain rule?

Multivariate Setting

- Conditional distribution, $p(y \mid x) = \dfrac{p(x, y)}{p(x)}$, Marginal $p(y) = \sum_{x \in \mathcal{X}} p(x, y)$
- Chain Rule $p(x, y) = p(y \mid x)p(x) = p(x \mid y)p(y)$
- Bayes Rule $p(y \mid x) = \dfrac{p(x \mid y)p(y)}{p(x)}$
- Law of total probability $p(y) = \sum_{x \in \mathcal{X}} p(y|x)p(x)$
- **Question**: How do you get the law of total probability from the chain rule?

$$p(y) = \sum_{x \in \mathcal{X}} p(x, y) = \sum_{x \in \mathcal{X}} p(y|x)p(x)$$

Question

- Assume $X \in \{0,1\}$ and $p(y \mid X = x)$ is Gaussian
    - We have $p(y \mid X = 0)$ is $\mathcal{N}(\mu_0, \sigma_0^2)$ and $p(y \mid X = 1)$ is $\mathcal{N}(\mu_1, \sigma_1^2)$
- Does this mean Y is Gaussian? (i.e., $p(y)$ is a Gaussian pdf)

Question

- Assume $X \in \{0,1\}$ and $p(y \mid X = x)$ is Gaussian
  - We have $p(y \mid X = 0)$ is $\mathcal{N}(\mu_0, \sigma_0^2)$ and $p(y \mid X = 1)$ is $\mathcal{N}(\mu_1, \sigma_1^2)$
- Does this mean Y is Gaussian? (i.e., $p(y)$ is a Gaussian pdf)
- No. In fact, it is a mixture of two Gaussians (like in your assignment) $p(y) = p(y \mid X = 0)p(X = 0) + p(y \mid X = 1)p(X = 1) = c_0\mathcal{N}(\mu_0, \sigma_0^2) + c_1\mathcal{N}(\mu_1, \sigma_1^2)$
- You did not need to know it is a mixture of Gaussians, but you should know that the conditional distribution over an RV and its marginals are not necessarily the same type of distribution; conditioning on more information results in a different distribution over Y (typically a lower variance one)

Expectations

$$\mathbb{E}[f(X)] = \begin{cases} \sum_{x \in \mathcal{X}} f(x) p(x) & \text{if } X \text{ is discrete,} \\ \int_{\mathcal{X}} f(x) p(x)\, dy & \text{if } X \text{ is continuous.} \end{cases}$$

Expectations

$$\mathbb{E}[f(X)] = \begin{cases} \sum_{x \in \mathcal{X}} f(x)p(x) & \text{if } X \text{ is discrete,} \\ \int_{\mathcal{X}} f(x)p(x)\, dy & \text{if } X \text{ is continuous.} \end{cases}$$

Eg: $\mathcal{X} = \{1,2,3,4,5\}$, $f(x) = x^2$, $Y = f(X)$, map $\{1,2,3,4,5\} \rightarrow \{1,4,9,16,25\}$, $p(y)$ determined by $p(x)$, e.g, $p(Y = 4) = p(X = 2)$

Expectations

$$\mathbb{E}[f(X)] = \begin{cases} \sum_{x\in\mathcal{X}} f(x)p(x) & \text{if } X \text{ is discrete,} \\ \int_{\mathcal{X}} f(x)p(x)\,dy & \text{if } X \text{ is continuous.} \end{cases}$$

Eg: $\mathcal{X} = \{1,2,3,4,5\}$, $f(x) = x^2$, $Y = f(X)$, map $\{1,2,3,4,5\} \rightarrow \{1,4,9,16,25\}$, $p(y)$ determined by $p(x)$, e.g, $p(Y=4) = p(X=2)$

Eg: $\mathcal{X} = \{-1,0,1\}$, $f(x) = |x|$, $Y = f(X)$, map $\{-1,0,1\} \rightarrow \{0,1\}$ $p(Y=1) = p(X=-1) + p(X=1)$, $\mathbb{E}[Y] = \sum_{y\in 0,1} yp(y) = \sum_{x\in\{-1,0,1\}} f(x)p(x)$

Conditional Expectations

Definition:

The **expected value of Y conditional on $X = x$** is

$$\mathbb{E}[Y \mid X = x] = \begin{cases} \sum_{y \in \mathcal{Y}} y p(y \mid x) & \text{if } Y \text{ is discrete,} \\ \int_{\mathcal{Y}} y p(y \mid x)\, dy & \text{if } Y \text{ is continuous.} \end{cases}$$

Conditional Expectations

Definition:

The **expected value of Y conditional on $X = x$** is

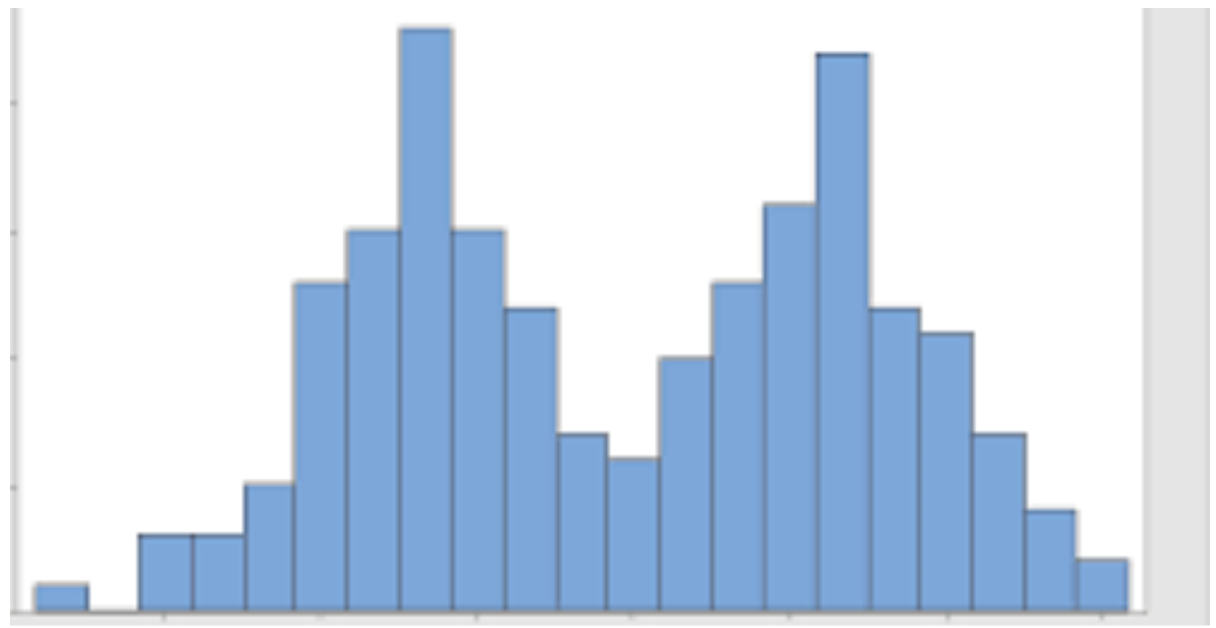
$$\mathbb{E}[Y \mid X = x] = \begin{cases} \sum_{y \in \mathcal{Y}} y p(y \mid x) & \text{if } Y \text{ is discrete,} \\ \int_{\mathcal{Y}} y p(y \mid x)\, dy & \text{if } Y \text{ is continuous.} \end{cases}$$

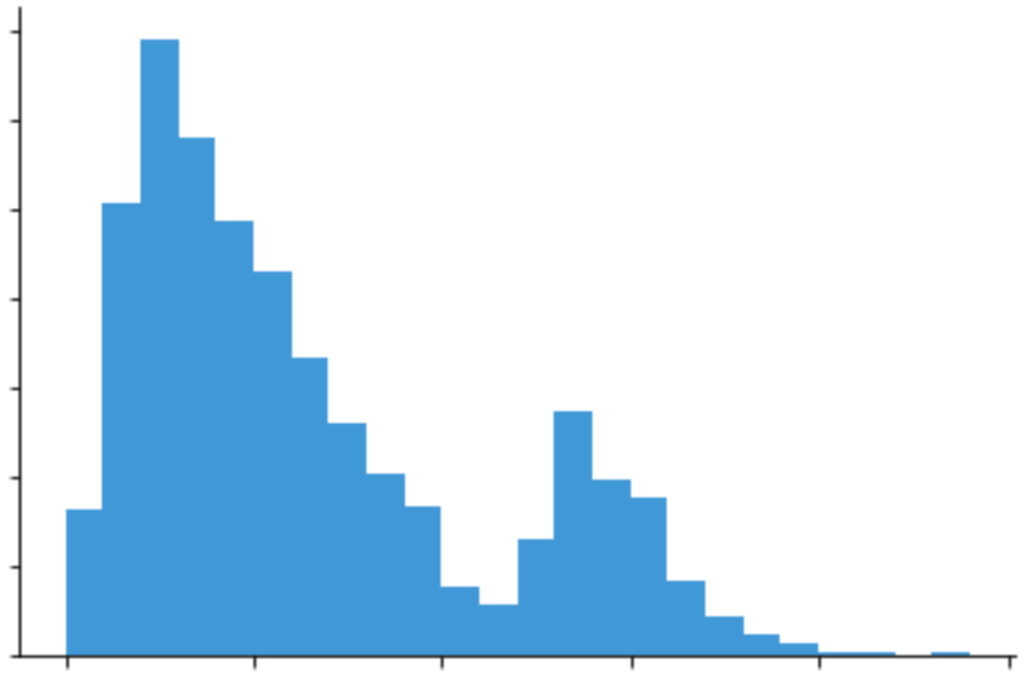
Recall Conditional Expectation Example

- X is the type of a book, 0 for fiction and 1 for non-fiction
  - $p(X = 1)$ is the proportion of all books that are non-fiction
- Y is the number of pages
  - $p(Y = 100)$ is the proportion of all books with 100 pages
- $p(y \mid X = 0)$ is different from $p(y \mid X = 1)$
- $\mathbb{E}[Y \mid X = 0]$ is different from $\mathbb{E}[Y \mid X = 1]$
  - e.g. $\mathbb{E}[Y \mid X = 0] = 70$ is different from $\mathbb{E}[Y \mid X = 1] = 150$

Conditional Expectation Example (cont)

- $p(y \mid X = 0)$ $p(y \mid X = 1)$





- $\mathbb{E}[Y \mid X = 0]$ is the expectation over Y under distribution $p(y \mid X = 0)$
- $\mathbb{E}[Y \mid X = 1]$ is the expectation over Y under distribution $p(y \mid X = 1)$

What if Y is dollars earned?

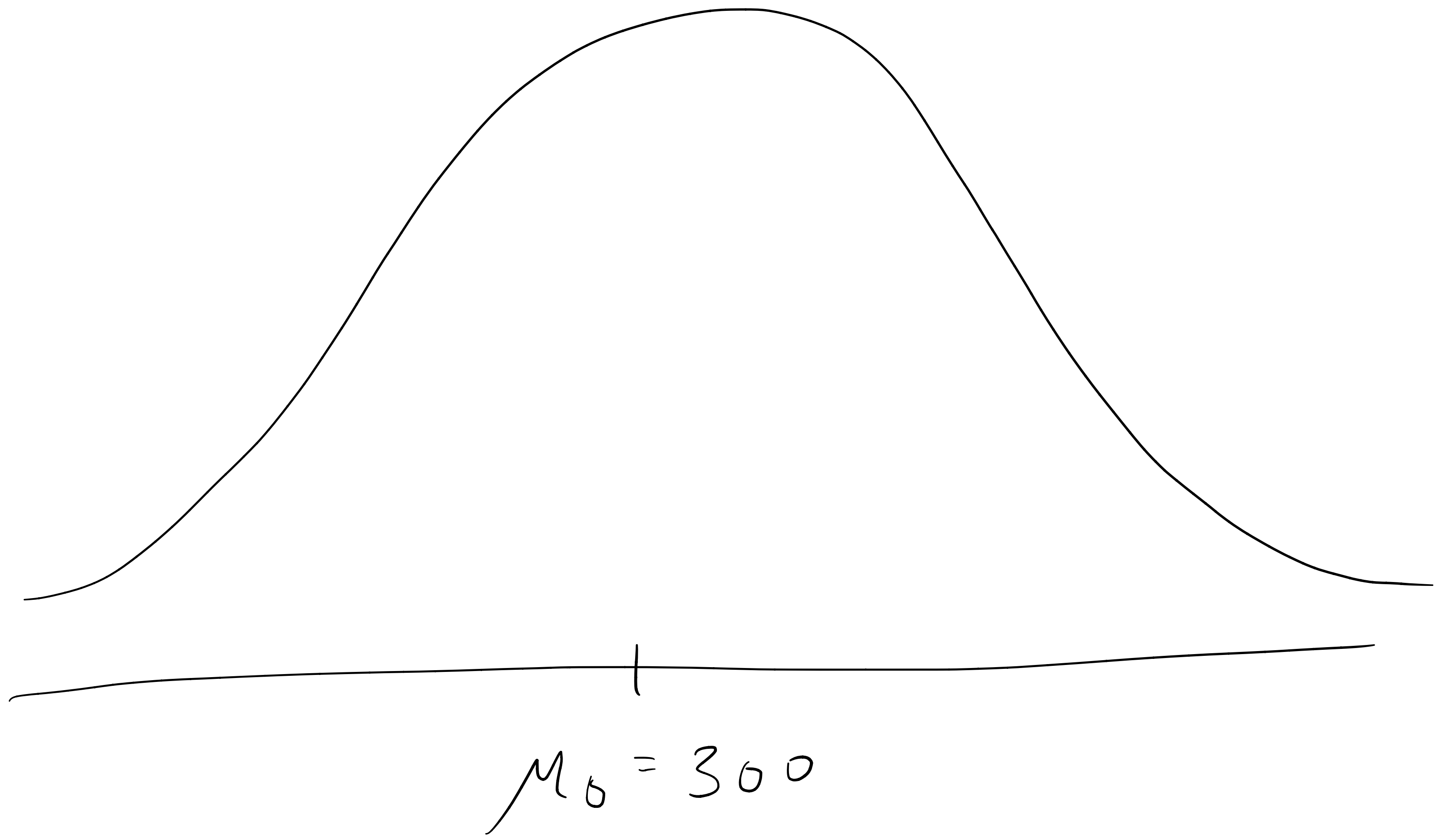
- Y is now a continuous RV
- Notice that $p(y|x)$ is defined by $p(y|X=0)$ and $p(y|X=1)$
- What might be a reasonable choice for $p(y|X=0)$ and $p(y|X=1)$?

What if Y is dollars earned?

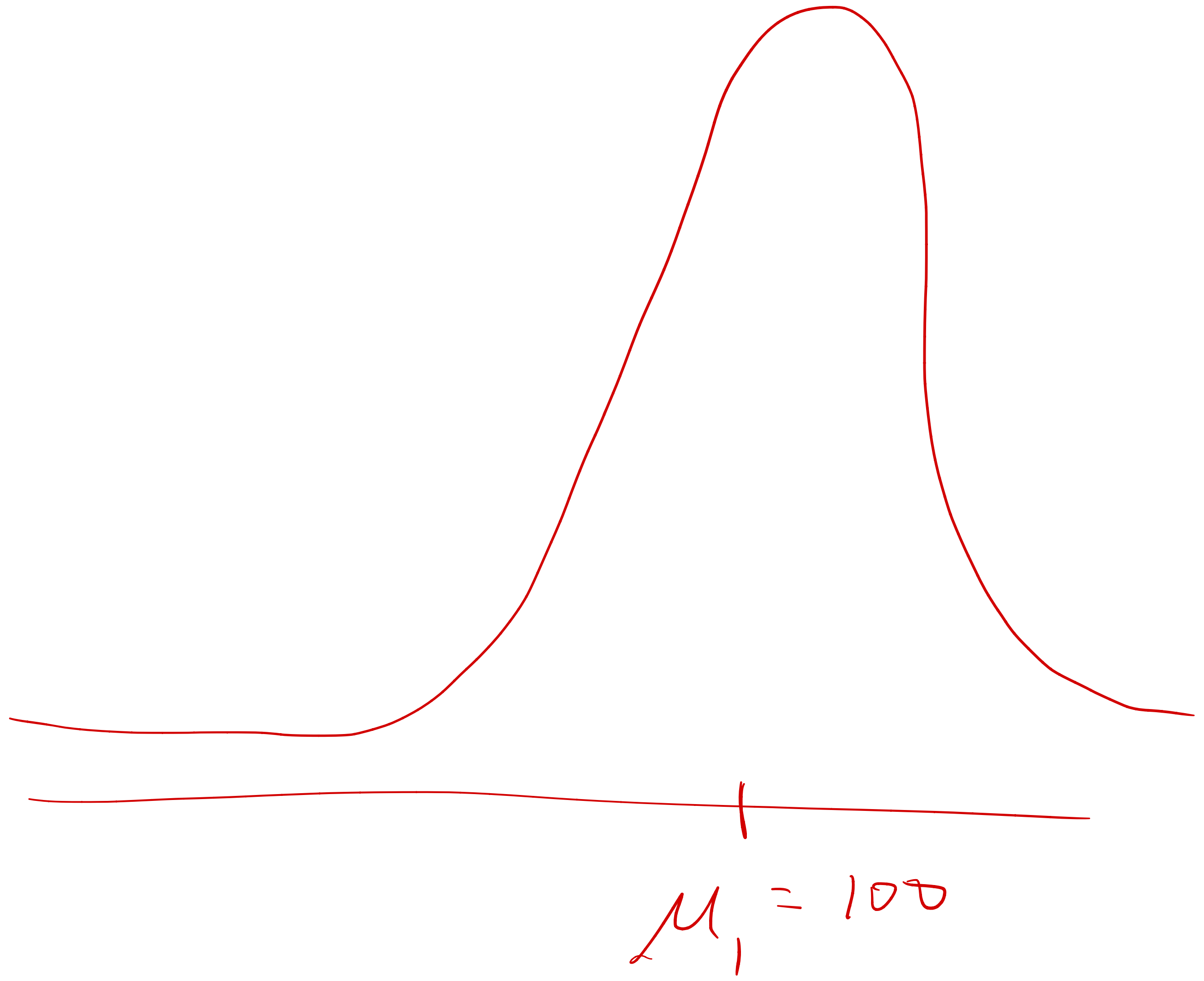
- Notice that $p(y \mid x)$ is defined by $p(y \mid X = 0)$ and $p(y \mid X = 1)$

$p(y \mid X=0) = \mathcal{N}(\mu_0, \sigma_0^2)$

$p(y \mid X=1) = \mathcal{N}(\mu_1, \sigma_1^2)$



Non-fiction



Fiction

Exercises

- Come up with an example of X and Y, and give possible choices for p(y | x)
- Do you need to know p(x) to use p(y | x)?
- If Y is discrete, then does X have to be discrete to specify p(y | x)?
- If we have p(y | x), can we get p(x | y)? Why or why not?

Exercises

- Do you need to know p(x) to use p(y | x)? **No.** If I want p(y | x =20) for x temperature and y humidity, I do not need to know p(x = 20)
- If Y is discrete, then does X have to be discrete to specify p(y | x)?
  - **No**. Y and X can be of different types (as we say with the books example).
  - Note: if X is continuous, we can ask p(y | x), because we are not asking Probability of x (which is zero), but rather defining the pdf/pmf over Y when conditioning on the fact that we observed x happening
- If we have p(y | x), can we get p(x | y)? Why or why not? **No**, we also need p(x) and p(y), and then we can use Bayes rule.

Properties of Expectations

- Linearity of expectation:
  - $\mathbb{E}[cX] = c\mathbb{E}[X]$ for all constant c
  - $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- Products of expectations of **independent** random variables X, Y:
  - $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

You should know linearity of expectation

Variance

Definition: The **variance** of a random variable is

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

i.e., $\mathbb{E}[f(X)]$ where $f(x) = (x - \mathbb{E}[X])^2$.

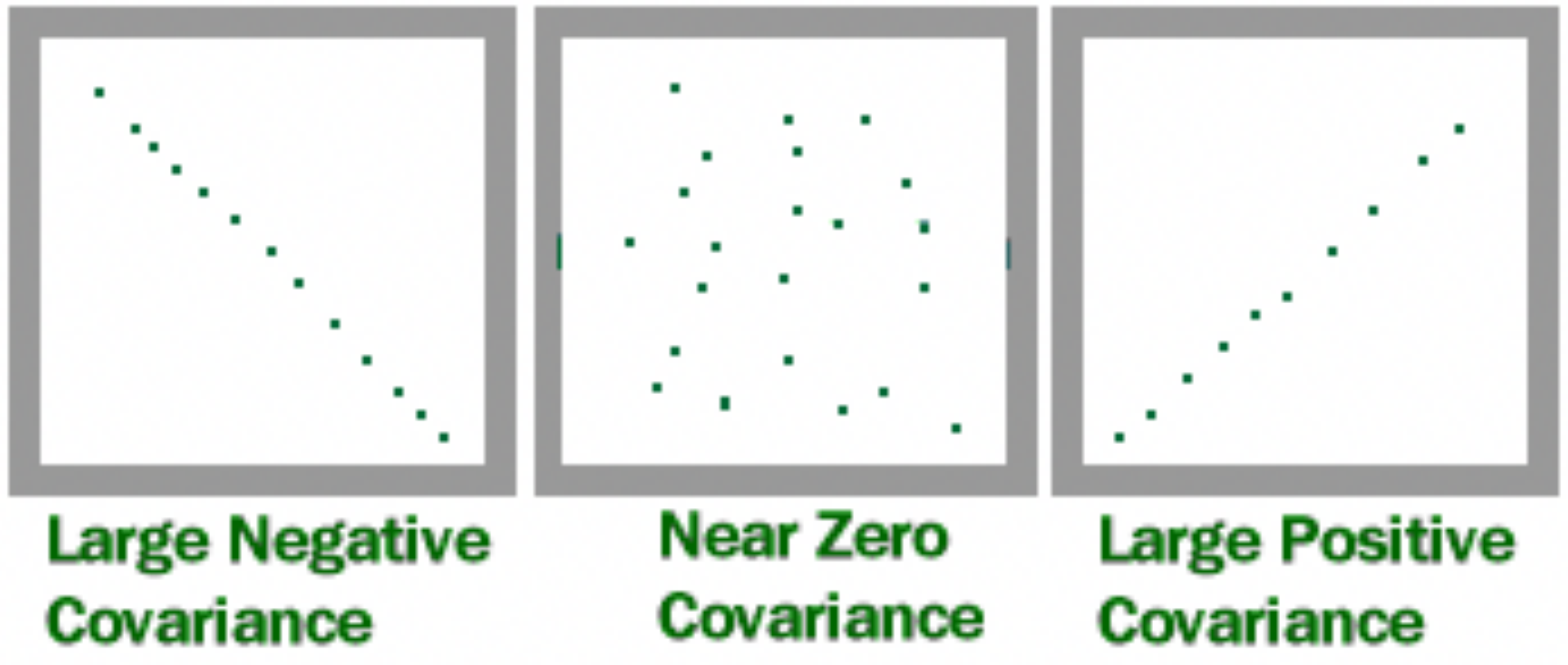
Equivalently,

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

Covariance

Definition: The covariance of two random variables is

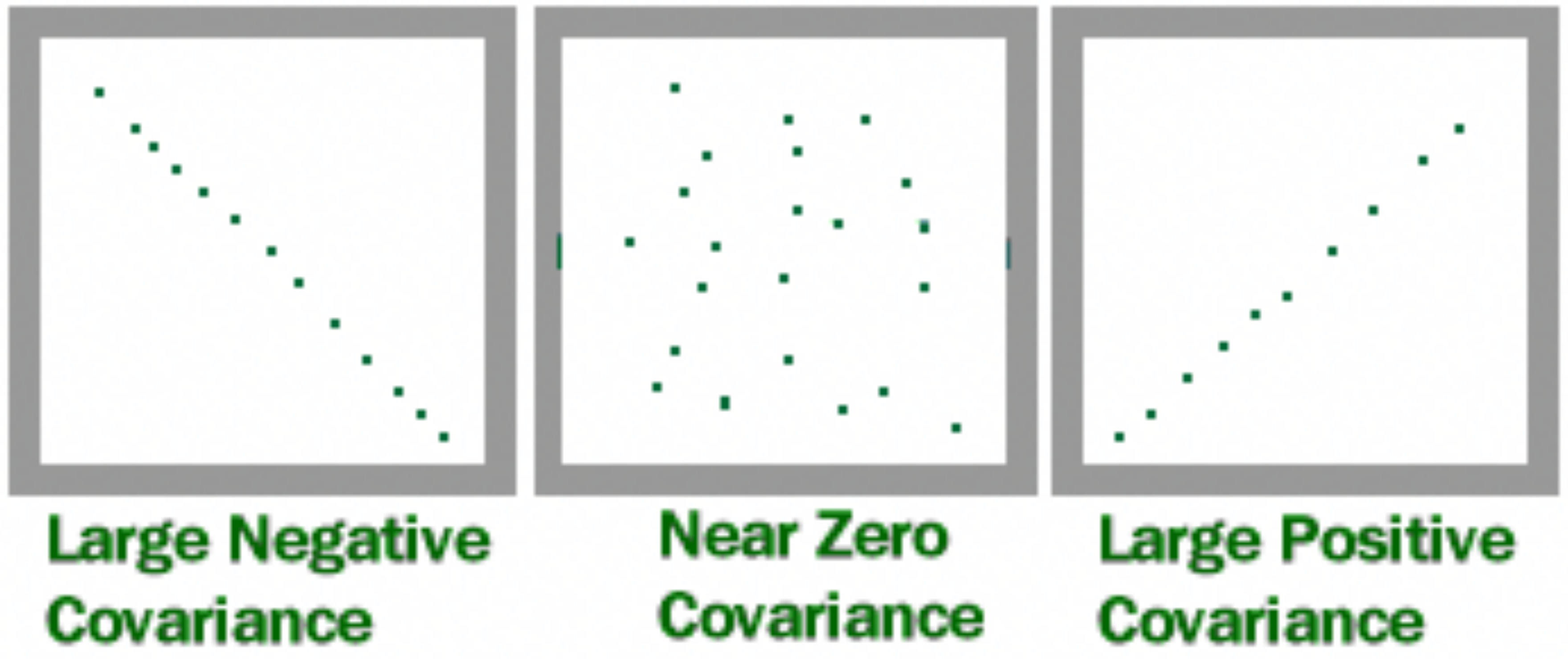
$$\begin{aligned}\text{Cov}(X,Y) &= \mathbb{E}\left[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\right] \\ &= \mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]\,.\end{aligned}$$



Covariance

Definition: The covariance of two random variables is

$$\text{Cov}(X,Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$
$$= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\,.$$



Properties of Variances

You should know all these properties

- Var$[c] = 0$ for constant c
- Var$[cX] = c^2$Var$[X]$ for constant c
- Var$[X + Y] =$ Var$[X] +$ Var$[Y] + 2$Cov$[X, Y]$
- For **independent** X, Y, because Cov$[X, Y] = 0$
  Var$[X + Y] =$ Var$[X] +$ Var$[Y]$

Let Y = 2X. What is Var(X + Y)?

Let $X = X_1, Y = X_2$ for iid samples X_1, X_2. What is Var(X + Y)?

Properties of Variances

Let Y = 2X. What is Var(X + Y)?
Option 1: Var(X+Y) = Var(X) + Var(2X) + 2 Cov(X,2X)
= Var(X) + 4Var(X) + 4 Var(X) = 9 Var(X)
Option 2: Var(X+Y) = Var(3X) = 9 Var(X)

- $\text{Var}[c] = 0$ for constant c
- $\text{Var}[cX] = c^2\text{Var}[X]$ for constant c
- $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$
- For **independent** X, Y, because $\text{Cov}[X, Y] = 0$
  $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$

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**

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

Properties of Variances (cont)

- Var$[c] = 0$ for constant c
- Var$[cX] = c^2$Var$[X]$ for constant c
- Var$[X + Y] =$ Var$[X] +$ Var$[Y] + 2$Cov$[X, Y]$
- For **independent** X, Y, because Cov$[X, Y] = 0$
  Var$[X + Y] =$ Var$[X] +$ Var$[Y]$

Let Y = 2X. What is Var(X + Y)?
Option 1: Var(X+Y) = Var(X) + Var(2X) + 2 Cov(X,2X)
= Var(X) + 4Var(X) + 4 Var(X) = 9 Var(X)
Option 2: Var(X+Y) = Var(3X) = 9 Var(X)

Let $X = X_1, Y = X_2$ for iid samples X_1, X_2. Let σ^2 be variance for X_1, X_2.
What is Var(X + Y)?

$$\text{Var}(X + Y) = \text{Var}(X_1) + \text{Var}(X_2) + 2\text{Cov}(X_1, X_2)$$

$$= \text{Var}(X_1) + \text{Var}(X_2) = 2\sigma^2$$

Estimators

Estimating Expected Value via the Sample Mean

We have n i.i.d. samples from the same distribution p, 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.

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

Bias

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

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

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

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.

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

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** shrinks
- We don't necessarily care about an estimator 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 μ:

$$\text{MSE}(\hat{X}) = \mathbb{E}\left[(\hat{X} - \mu)^2\right] \text{ where } \mathbb{E}[\hat{X}] \text{ may not equal } \mu$$

Bias-Variance Tradeoff

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

- If we can decrease variance without increasing bias, error goes down
- Biasing the estimator toward values that are **more likely to be true** based on **prior information**

Bias-Variance Tradeoff

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

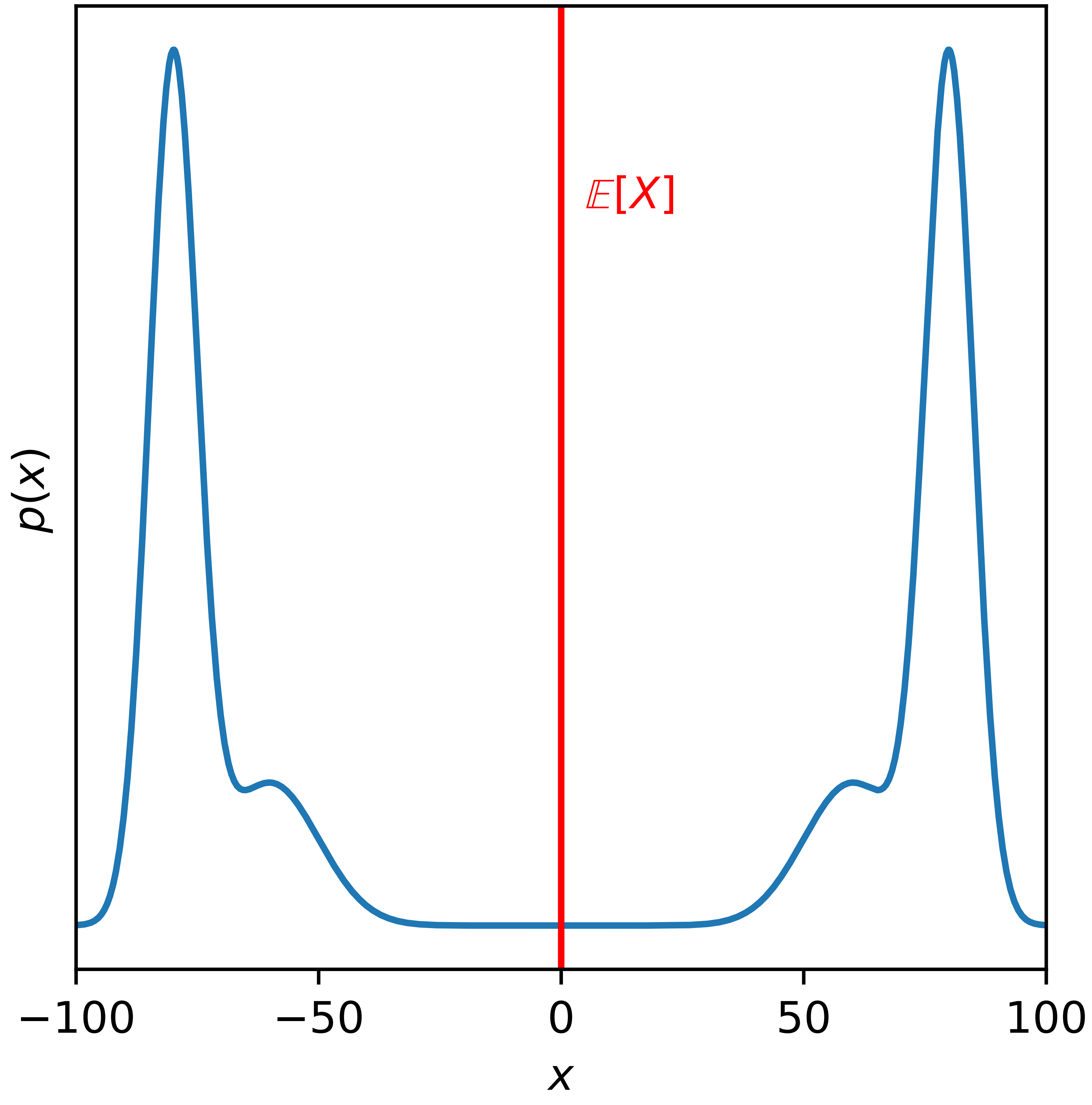
- Biasing the estimator toward values that are **more likely to be true** based on **prior information**
- Example: over five years you have computed that a typical average number of accidents $k = 5$ for factories of a medium size
- You want to estimate the average number of accidents for a new factory, but only have a weeks worth of data
- A reasonable (biased) estimator is: $\frac{1}{8}[k + \sum_{i=1}^{7} x_i]$

  Or an even lower variance (higher bias) is $\frac{1}{10}[3k + \sum_{i=1}^{7} x_i]$

Why is bias higher?

- Imagine $k = 5$ and the true mean is $\mu = 4$
- $\mathbb{E}\left[\frac{1}{8}(k + \sum_{i=1}^{7} X_i)\right] = \frac{1}{8}\left(k + \mathbb{E}[\sum_{i=1}^{7} X_i]\right) = \frac{1}{8}\left(k + 7\mu\right) = \frac{1}{8}\left(5 + 7 \times 4\right) = \frac{33}{8} = 4.13 \neq 4$
- $\mathbb{E}\left[\frac{1}{10}(3k + \sum_{i=1}^{7} X_i)\right] = \frac{1}{10}\left(3k + \mathbb{E}[\sum_{i=1}^{7} X_i]\right) = \frac{1}{10}\left(3k + 7\mu\right) = \frac{1}{10}\left(3 \times 5 + 7 \times 4\right) = \frac{43}{10} = 4.3 \neq 4$
- You can check that the variance is slightly lower for the second one, since it is like it has 10 samples instead of 8, and both are lower than the unbiased sample mean

Prior information helps overcome high variance in sampling



It’s possible in a small sample to see only data between 50 and 100

The sample mean is highly inaccurate
due to the high variance in this distribution

Once we have lots of data, this problem disappears
We really only care about introducing bias to reduce variance for smaller sample sizes

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 $\mu = 0.1$

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

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

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

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

Exercise: What is the variance of these estimators?

Questions:

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 $Z \sim N(0,100^2)$
3. $Y = \frac{1}{n}\sum_{i=1}^{n} X_i$, for $X_i \sim p$

Exercise: What is the variance of these estimators?

Estimators:

1. $Y_1 \sim$ Uniform$[0,10]$
2. $Y_2 = \mathbb{E}[X] + Z,$ where $Z \sim N(0,100^2)$
3. $Y_3 = \frac{1}{n}\sum_{i=1}^{n} X_i$, for $X_i \sim p$

$$\text{Var}(Y_1) = \frac{1}{12}(10-0)^2 = \frac{100}{12} = 8.\bar{3}$$

$$\text{Var}(Y_2) = \text{Var}(\mathbb{E}[X] + Z) = ?$$

$$\text{Var}(Y_3) = \frac{\sigma^2}{n}$$

Exercise: What is the variance of these estimators?

Estimators:

1. $Y_1 \sim$ Uniform$[0,10]$
2. $Y_2 = \mathbb{E}[X] + Z,$ where $Z \sim N(0,100^2)$
3. $Y_3 = \frac{1}{n}\sum_{i=1}^{n} X_i$, for $X_i \sim p$

$$\begin{aligned} \text{Var}(Y_2) &= \text{Var}(\mathbb{E}[X] + Z) \\ &= \text{Var}(Z) && \triangleright\ \text{Var}(c + Y) = \text{Var}(Y) \\ &= 100^2 \end{aligned}$$

MSE of these estimators

$$\text{Var}(Y_1) = \frac{1}{12}(10-0)^2 = \frac{100}{12} = 8.\bar{3}$$

$$\text{Bias}(Y_1) = \mathbb{E}[Y_1] - \mathbb{E}[X] = 5$$

$$\text{Var}(Y_2) = \text{Var}(\mathbb{E}[X] + Z) = 100^2$$

$$\text{Bias}(Y_2) = \mathbb{E}[Y_2] - \mathbb{E}[X] = 0$$

$$\text{Var}(Y_3) = \frac{\sigma^2}{n}$$

$$\text{Bias}(Y_3) = 0$$

Estimators:

1. $Y_1 \sim$ Uniform$[0,10]$
2. $Y_2 = \mathbb{E}[X] + Z,$ where $Z \sim N(0,100^2)$
3. $Y_3 = \frac{1}{n}\sum_{i=1}^{n} X_i,$ for $X_i \sim p$

$$\text{MSE}(Y_1) = 5^2 + 8.\bar{3} = 33.\bar{3}$$

$$\text{MSE}(Y_2) = 0 + 100^2 = 10000$$

$$\text{MSE}(Y_3) = 0 + \frac{\sigma^2}{n}$$

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

Concentration Inequalities

- We would like to be able to claim $\Pr\left(\left|\bar{X}-\mu\right| < \epsilon\right) > 1-\delta$ for some $\delta, \epsilon > 0$

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, for $\delta \in (0,1)$, $\Pr\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \leq (b-a)\sqrt{\frac{\ln(2/\delta)}{2n}}\right) \geq 1 - \delta.$

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, for $\delta \in (0,1)$, $\Pr\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \leq \sqrt{\frac{\sigma^2}{\delta n}}\right) \geq 1 - \delta.$

When to Use Chebyshev, When to Use Hoeffding?

- If $a \leq X_i \leq b$, then $\text{Var}[X_i] \leq \frac{1}{4}(b-a)^2$
- 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 $\epsilon = \sqrt{\frac{\sigma^2}{\delta n}} \leq \sqrt{\frac{(b-a)^2}{4\delta n}} = \frac{1}{2\sqrt{\delta}}(b-a)\sqrt{\frac{1}{n}}$
- **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

Sample Complexity

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

- We want sample complexity to be small
- Sample complexity is determined by:
  1. The **estimator** itself
     - Smarter estimators can sometimes improve sample complexity (e.g., smart priors)
  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

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

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

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

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

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

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

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

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

Summary

- **Concentration inequalities** let us bound the probability of a given estimator being at least ϵ from its mean (expected value)
- **Sample complexity** is the **number of samples** needed to attain a desired error bound ϵ at a desired probability $1 - \delta$
  - We only discussed sample complexity for unbiased estimators
- 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 on some **prior information**
  - *But this only helps if the prior information is **correct,** cannot reduce error by adding in arbitrary bias*

Optimization

- Represent a problem as an optimization problem
- Solve a discrete problem by iterating over options and picking the one with the minimum value according to the objective
- Solve a continuous optimization problem by finding **stationary points**
  - A point w is a stationary point if $c'(w) = 0$
  - or for multivariate $\mathbf{w}$, $\nabla c(\mathbf{w}) = 0$

Poll Question: Which of the following are true about stationary points?

- 1. A stationary point is the global minimum of a function
- 2. A stationary point is a point where the gradient is zero
- 3. A global minimum is a stationary point, but a stationary point may not be a global minimum
- 4. If we find a stationary point, then we have found the minimum of our function
- 5. We can use the second derivative test to identify the type of stationary point we have

Poll Question: Which of the following are true about stationary points?

- 1. A stationary point is the global minimum of a function
- 2. A stationary point is a point where the gradient is zero
- 3. A global minimum is a stationary point, but a stationary point may not be a global minimum
- 4. If we find a stationary point, then we have found the minimum of our function
- 5. We can use the second derivative test to identify the type of stationary point we have

Answer: 2, 3 and 5

Optimization

- Represent a problem as an optimization problem
- Solve an optimization problem by finding **stationary points**
- **Define first-order gradient descent**
- **Define second-order gradient descent**
- Define **step size** and **adaptive step size**
- Explain the role and importance of step sizes in first-order gradient descent
- Apply gradient descent to numerically find local optima

Exercise

- Imagine $c(w) = \frac{1}{2}(xw - y)^2$.
- What is the first-order update, assuming we are currently at point w_t?
    - i.e., the gradient descent update tells us how to modify our current point to descend on our surface c.

Exercise

- Imagine $c(w) = \frac{1}{2}(xw - y)^2$.
- What is the first-order update, assuming we are currently at point w_t?
  - i.e., the gradient descent update tells us how to modify our current point to descend on our surface c.

Answer: $w_{t+1} \leftarrow w_t - \eta_t c'(w_t)$ for some stepsize $\eta_t > 0$

$c'(w) = (xw - y)x$ so we have that. $w_{t+1} \leftarrow w_t - \eta_t(xw_t - y)x$

Exercise

- Imagine $c(w) = \frac{1}{2}(xw - y)^2$.
- What is the first-order update, assuming we are currently at point w_t?
  - i.e., the gradient descent update tells us how to modify our current point to descend on our surface c.
- What if instead we did $w_{t+1} \leftarrow w_t + \eta_t c'(w_t)$. What would happen?
- The second-order update is $w_{t+1} \leftarrow w_t - \dfrac{c'(w_t)}{c''(w_t)}$. Why might this update be preferable to the first-order? (poll)

Poll Question: Why might the second-order update be preferable?

- 1. It is easier to compute than the first-order one.
- 2. It tells us how to pick a good stepsize.
- 3. The second-order update is more likely to get stuck at a saddlepoint
- 4. The first-order update might get stuck in local minimum, but not the second-order update

Poll Question: Why might the second-order update be preferable?

- 1. It is easier to compute than the first-order one.
- 2. It tells us how to pick a good stepsize.
- 3. The second-order update is more likely to get stuck at a saddlepoint
- 4. The first-order update might get stuck in local minimum, but not the second-order update

Answer: 2

Closed-form solutions

- $c(w) = (w-3)^2$ has a closed-form solution because

$$c'(w) = 2(w-3) = 0 \implies w - 3 = 0 \implies w = 3.$$

- $c(w) = w^2 + \exp(w)$ does not have a closed-form solution because

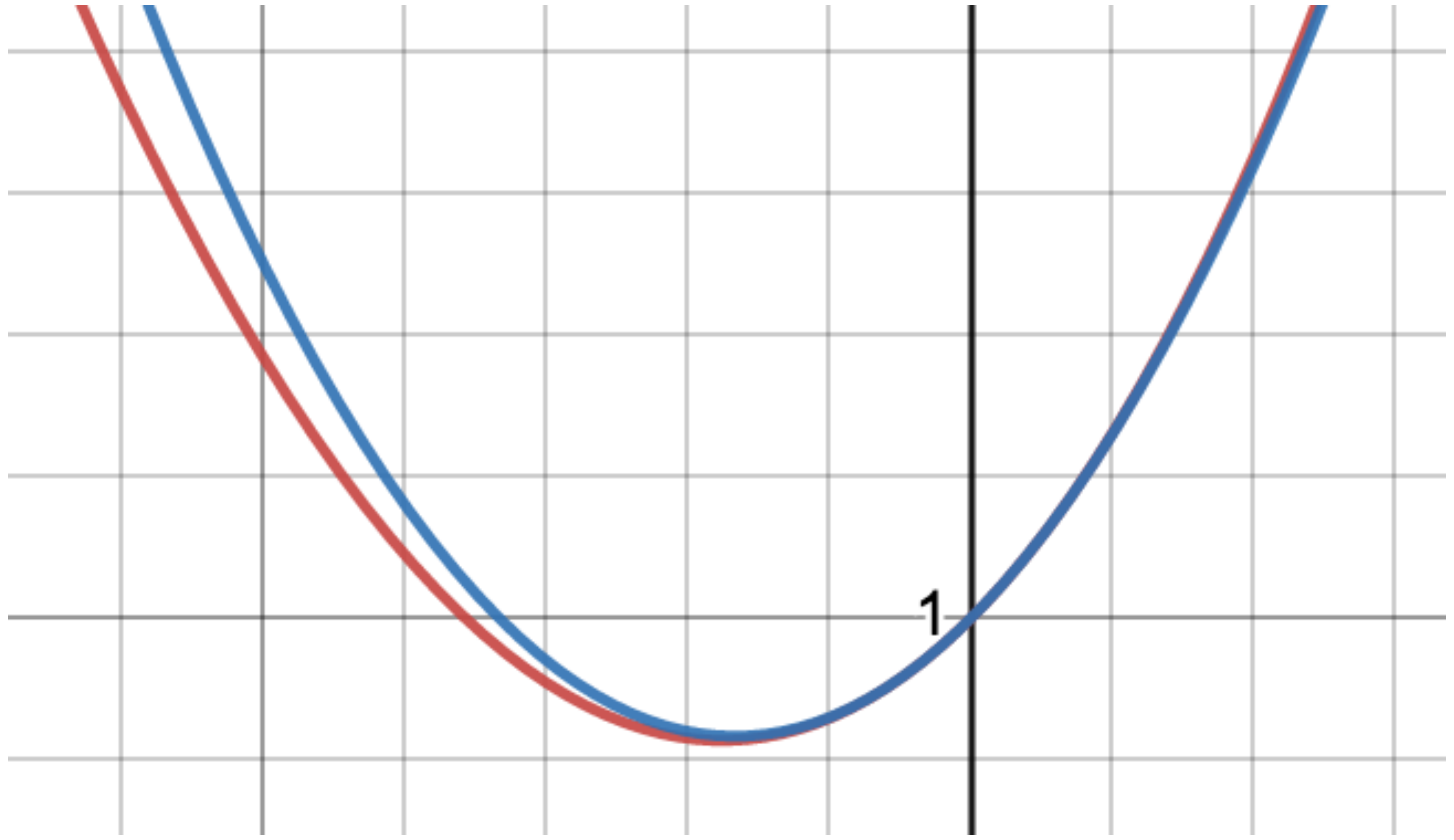
$$c'(w) = 2w + \exp(w) = 0 \implies \exp(w) = -2w$$

- Can’t isolate w on one side, to get an explicit formula (closed-form)
  - Note: this c is not a hard optimization problem, it is convex

Second-order update

Example 14: Let us revisit our example $c(w) = w^2 + \exp(w)$, where $c'(w) = 2w + \exp(w)$ and $c''(w) = 2 + \exp(w)$. Let us start $w_0 = 0$ and do one second-order update.

$$\begin{aligned} w_1 &= w_0 - \frac{c'(w_0)}{c''(w_0)} \\ &= 0 - \frac{0 + \exp(0)}{2 + \exp(0)} \\ &= -\frac{1}{3} \end{aligned}$$



red line is c(w),
blue line is second-order Taylor approximation
around w = 0

Now let us do the next update.

$$\begin{aligned} w_2 &= w_1 - \frac{c'(w_1)}{c''(w_1)} \\ &= -\frac{1}{3} - \frac{-\frac{2}{3} + \exp(-\frac{1}{3})}{2 + \exp(-\frac{1}{3})} \\ &= -0.3516893316 \end{aligned}$$

$$\begin{aligned} \hat{c}(w) &= c(w_0) + (w - w_0)c'(w_0) + \frac{1}{2}(w - w_0)^2 c''(w_0) \\ &= \exp(0) + w\exp(0) + (2 + \exp(0))\frac{1}{2}w^2 = 1 + w + \tfrac{3}{2}w^2 \end{aligned}$$

Things you do not need to know for the exam

- You do not need to know the formulas for any pdfs or pmfs
- You should be comfortable with Bayes rule, chain rule for probability and expectation/variance rules, though I will typically remind you of these rules
- You should know basic math rules, like $\ln \exp(a) = a$
- You do not have to remember the Chebyshev's or Hoeffding's inequality, but you do have to know how to use them
- You will not have to compute any derivatives or integrals