Probability, continued

CMPUT 296: Basics of Machine Learning

§2.2-2.4
Recap

• Probabilities are a means of quantifying uncertainty

• A probability distribution is defined on a measurable space consisting of a sample space and an event space.

• **Discrete** sample spaces (and random variables) are defined in terms of probability mass functions (PMFs)

• **Continuous** sample spaces (and random variables) are defined in terms of probability density functions (PDFs)
Outline

1. Multiple Random Variables
2. Independence
3. Expectations and Moments
Recap: Random Variables

**Random variables** are a way of reasoning about a complicated underlying probability space in a more straightforward way.

**Example:** Suppose we observe both a die's number, and where it lands.

\[ \Omega = \{(left,1), (right,1), (left,2), (right,2), \ldots, (right,6)\} \]

We might want to think about the probability that we get a large number, without thinking about where it landed.

We could ask about \( P(X \geq 4) \), where \( X = \text{number that comes up} \).
What About Multiple Variables?

- So far, we've really been thinking about a single random variable at a time
- Straightforward to define multiple random variables on a single probability space

Example: Suppose we observe both a die's number, and where it lands.

\[ \Omega = \{(left,1), (right,1), (left,2), (right,2), \ldots, (right,6)\} \]

\[ X(\omega) = \omega_2 = \text{number} \]

\[ Y(\omega) = \begin{cases} 
  1 & \text{if } \omega_1 = \text{left} \\
  0 & \text{otherwise.}
\end{cases} \]

\[ P(Y = 1) = P(\{\omega \mid Y(\omega) = 1\}) \]

\[ P(X \geq 4 \land Y = 1) = P(\{\omega \mid X(\omega) \geq 4 \land Y(\omega) = 1\}) \]
Joint Distribution

We typically model the interactions of different random variables.

**Joint probability mass function:** \( p(x, y) = P(X = x, Y = y) \)

\[
\sum_{x \in X} \sum_{y \in Y} p(x, y) = 1
\]

**Example:** \( X = \{0,1\} \) (young, old) and \( Y = \{0,1\} \) (no arthritis, arthritis)

<table>
<thead>
<tr>
<th></th>
<th>( Y=0 )</th>
<th>( Y=1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X=0 )</td>
<td>( P(X=0,Y=0) = \frac{1}{2} )</td>
<td>( P(X=0, Y=1) = \frac{1}{100} )</td>
</tr>
<tr>
<td>( X=1 )</td>
<td>( P(X=1, Y=0) = \frac{1}{10} )</td>
<td>( P(X=1, Y=1) = \frac{39}{100} )</td>
</tr>
</tbody>
</table>
Questions About Multiple Variables

Example: $\mathcal{X} = \{0, 1\}$ (young, old) and $\mathcal{Y} = \{0, 1\}$ (no arthritis, arthritis)

<table>
<thead>
<tr>
<th></th>
<th>$Y=0$</th>
<th>$Y=1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X=0$</td>
<td>$P(X=0, Y=0) = \frac{1}{2}$</td>
<td>$P(X=0, Y=1) = \frac{1}{100}$</td>
</tr>
<tr>
<td>$X=1$</td>
<td>$P(X=1, Y=0) = \frac{1}{10}$</td>
<td>$P(X=1, Y=1) = \frac{39}{100}$</td>
</tr>
</tbody>
</table>

- Are these two variables related at all? Or do they change independently?
- Given this distribution, can we determine the distribution over just $Y$? I.e., what is $P(Y = 1)$? (marginal distribution)
- If we knew something about one variable, does that tell us something about the distribution over the other? E.g., if I know $X = 0$ (person is young), does that tell me the conditional probability $P(Y = 1 \mid X = 1)$? (Prob. that person we know is young has arthritis)
Conditional Distribution

**Definition:** Conditional probability distribution

\[
P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)}
\]

This same equation will hold for the corresponding PDF or PMF:

\[
p(y \mid x) = \frac{p(x, y)}{p(x)}
\]

**Question:** if \(p(x, y)\) is small, does that imply that \(p(y \mid x)\) is small?

e.g., imagine \(x = \text{arthritis}\) and \(y = \text{old}\)
PMFs and PDFs of Many Variables

In general, we can consider a $d$-dimensional random variable $\vec{X} = (X_1, \ldots, X_d)$ with vector-valued outcomes $\vec{x} = (x_1, \ldots, x_d)$, with each $x_i$ chosen from some $\mathcal{X}_i$. Then,

**Discrete case:**
$p : \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_d \to [0,1]$ is a (joint) probability mass function if

$$\sum_{x_1 \in \mathcal{X}_1} \sum_{x_2 \in \mathcal{X}_2} \cdots \sum_{x_d \in \mathcal{X}_d} p(x_1, x_2, \ldots, x_d) = 1$$

**Continuous case:**
$p : \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_d \to [0,\infty)$ is a (joint) probability density function if

$$\int_{\mathcal{X}_1} \int_{\mathcal{X}_2} \cdots \int_{\mathcal{X}_d} p(x_1, x_2, \ldots, x_d) \, dx_1 dx_2 \ldots dx_d = 1$$
A **marginal distribution** is defined for a subset of $\mathbf{X}$ by summing or integrating out the remaining variables. (We will often say that we are "marginalizing over" or "marginalizing out" the remaining variables).

**Discrete case:**

$$p(x_i) = \sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_{i-1} \in \mathcal{X}_{i-1}} \sum_{x_{i+1} \in \mathcal{X}_{i+1}} \cdots \sum_{x_d \in \mathcal{X}_d} p(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$$

**Continuous:**

$$p(x_i) = \int_{\mathcal{X}_1} \cdots \int_{\mathcal{X}_{i-1}} \int_{\mathcal{X}_{i+1}} \cdots \int_{\mathcal{X}_d} p(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d) \, dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_d$$
Back to our example

Example: \( \mathcal{X} = \{0,1\} \) (young, old) and \( \mathcal{Y} = \{0,1\} \) (no arthritis, arthritis)

<table>
<thead>
<tr>
<th></th>
<th>( Y=0 )</th>
<th>( Y=1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X=0 )</td>
<td>( P(X=0, Y=0) = 1/2 )</td>
<td>( P(X=0, Y=1) = 1/100 )</td>
</tr>
<tr>
<td>( X=1 )</td>
<td>( P(X=1, Y=0) = 1/10 )</td>
<td>( P(X=1, Y=1) = 39/100 )</td>
</tr>
</tbody>
</table>

**Exercise:** Check if \( \sum_{x \in \{0,1\}} \sum_{y \in \{0,1\}} p(x, y) = 1 \)

**Exercise:** Compute marginal \( p(y) = \sum_{x \in \{0,1\}} p(x, y) \)
Example: $\mathcal{X} = \{0,1\}$ (young, old) and $\mathcal{Y} = \{0,1\}$ (no arthritis, arthritis)

<table>
<thead>
<tr>
<th></th>
<th>$\mathcal{Y} = 0$</th>
<th>$\mathcal{Y} = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{X} = 0$</td>
<td>$P(X=0, Y=0) = 1/2$</td>
<td>$P(X=0, Y=1) = 1/100$</td>
</tr>
<tr>
<td>$\mathcal{X} = 1$</td>
<td>$P(X=1, Y=0) = 1/10$</td>
<td>$P(X=1, Y=1) = 39/100$</td>
</tr>
</tbody>
</table>

Exercise: Check if

\[
\sum_{x \in \{0,1\}} \sum_{y \in \{0,1\}} p(x, y) = \frac{1}{2} + \frac{1}{100} + \frac{1}{10} + \frac{39}{100} = 1
\]

Exercise: Compute marginal $p(y = 1) = \sum_{x \in \{0,1\}} p(x, y = 1) = \frac{40}{100}$,

\[
p(y = 0) = 1 - p(y = 1) = \frac{60}{100}
\]
A **marginal distribution** is defined for a subset of $\vec{X}$ by summing or integrating out the remaining variables. (We will often say that we are "marginalizing over" or "marginalizing out" the remaining variables).

**Discrete case:**

$$p(x_i) = \sum_{x_1 \in X_1} \cdots \sum_{x_{i-1} \in X_{i-1}} \sum_{x_{i+1} \in X_{i+1}} \cdots \sum_{x_d \in X_d} p(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$$

**Continuous:**

$$p(x_i) = \int_{X_1} \cdots \int_{X_{i-1}} \int_{X_{i+1}} \cdots \int_{X_d} p(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d) \, dx_1 \cdots dx_{i-1} \, dx_{i+1} \cdots dx_d$$

**Question:** How do we get $p(x_i, x_j)$ for some $i, j$?

**Question:** Why $p$ for $p(x_i)$ and $p(x_1, \ldots, x_d)$?

- They can't be the same function, they have different domains!
Are these really the same function?

- No. They're not the same function.
- But they are derived from the same joint distribution.
- So for brevity we will write

\[ p(y \mid x) = \frac{p(x, y)}{p(x)} \]

- Even though it would be more precise to write something like

\[ p_{Y\mid X}(y \mid x) = \frac{p(x, y)}{p_X(x)} \]

- We can tell which function we’re talking about from context (i.e., arguments)
Chain Rule

From the definition of conditional probability:

\[
p(y \mid x) = \frac{p(x, y)}{p(x)}
\]

\[\iff p(y \mid x)p(x) = \frac{p(x, y)}{p(x)}p(x)\]

\[\iff p(y \mid x)p(x) = p(x, y)\]

This is called the **Chain Rule**.
Multiple Variable Chain Rule

The chain rule generalizes to multiple variables:

\[
p(x, y, z) = p(x, y \mid z)p(z) = p(x \mid y, z)p(y \mid z)p(z)
\]

**Definition: Chain rule**

\[
p(x_1, \ldots, x_d) = p(x_d) \prod_{i=1}^{d-1} p(x_i \mid x_{i+1}, \ldots x_d)
\]

\[
= p(x_1) \prod_{i=2}^{d} p(x_i \mid x_i, \ldots x_{i-1})
\]
Bayes' Rule

From the chain rule, we have:

\[ p(x, y) = p(y \mid x)p(x) = p(x \mid y)p(y) \]

- Often, \( p(x \mid y) \) is easier to compute than \( p(y \mid x) \)
- e.g., where \( x \) is \textbf{features} and \( y \) is \textbf{label}

Definition: Bayes' rule

\[ p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)} \]

Posterior \quad Likelihood \quad Prior

Evidence
Example: Disease Test

Example:

\[ p(\text{Test} = \text{pos} \mid \text{Dis} = T) = 0.99 \]
\[ p(\text{Test} = \text{pos} \mid \text{Dis} = F) = 0.03 \]
\[ p(\text{Dis} = T) = 0.005 \]

Questions:

1. What is the likelihood?
2. What is the prior?
3. What is \( p(\text{Dis} = T \mid \text{Test} = \text{pos}) \)?
Independence of Random Variables

**Definition:** $X$ and $Y$ are **independent** if:

$$p(x, y) = p(x)p(y)$$

$X$ and $Y$ are **conditionally independent given $Z$** if:

$$p(x, y \mid z) = p(x \mid z)p(y \mid z)$$
Another Marginalization Example

• Imagine you get to draw two random candies from a bag of treats
• Say there are 5 types of candies (1, 2, 3, 4, 5), equally distributed in the bag
• Let $X =$ First Candy You Got and $Y =$ Second Candy You Got
• What is $p(X = 1)$?
• What is $p(X = 1, Y = 3)$?
Independence of Random Variables

**Definition:** $X$ and $Y$ are **independent** if:

$$p(x, y) = p(x)p(y)$$

$X$ and $Y$ are **conditionally independent given** $Z$ if:

$$p(x, y | z) = p(x | z)p(y | z)$$
Example: Coins
(Ex. 7 in the course text)

• Suppose you have a biased coin: It does not come up heads with probability 0.5. Instead, it is more likely to come up heads.

• Let $Z$ be the bias of the coin, with $\mathcal{Z} = \{0.3, 0.5, 0.8\}$ and probabilities
  $P(Z = 0.3) = 0.7$, $P(Z = 0.5) = 0.2$ and $P(Z = 0.8) = 0.1$.

• **Question:** What other outcome space could we consider?
• **Question:** What kind of distribution is this?
• **Question:** What other kinds of distribution could we consider?
Example: Coins (2)

- Now imagine I told you $Z = 0.3$ (i.e., probability of heads is 0.3)

- Let $X$ and $Y$ be two consecutive flips of the coin

- What is $P(X = \text{Heads} \mid Z = 0.3)$? What about $P(X = \text{Tails} \mid Z = 0.3)$?

- What is $P(Y = \text{Heads} \mid Z = 0.3)$? What about $P(Y = \text{Tails} \mid Z = 0.3)$?

- Is $P(X = x, Y = y \mid Z = 0.3) = P(X = x \mid Z = 0.3)P(Y = y \mid Z = 0.3)$?
Example: Coins (3)

• Now imagine we do not know $Z$
  
  • e.g., you randomly grabbed it from a bin of coins with probabilities
    \[ P(Z = 0.3) = 0.7, \ P(Z = 0.5) = 0.2 \text{ and } P(Z = 0.8) = 0.1 \]
  
  • What is $P(X = Heads)$?
    \[
P(X = Heads) = \sum_{z \in \{0.3, 0.5, 0.8\}} P(X = Heads | Z = z)p(Z = z)
    \]
    \[
    = P(X = Heads | Z = 0.3)p(Z = 0.3)
    + P(X = Heads | Z = 0.5)p(Z = 0.5)
    + P(X = Heads | Z = 0.8)p(Z = 0.8)
    \]
    \[
    = 0.3 \times 0.7 + 0.5 \times 0.2 + 0.8 \times 0.1 = 0.39
    \]
Now imagine we do not know $Z$

- e.g., you randomly grabbed it from a bin of coins with probabilities $P(Z = 0.3) = 0.7$, $P(Z = 0.5) = 0.2$ and $P(Z = 0.8) = 0.1$

- Is $P(X = \text{Heads}, Y = \text{Heads}) = P(X = \text{Heads})p(Y = \text{Heads})$?

- For brevity, let's use $h$ for Heads

$$P(X = h, Y = h) = \sum_{z \in \{0.3, 0.5, 0.8\}} P(X = h, Y = h \mid Z = z)p(Z = z)$$

$$= \sum_{z \in \{0.3, 0.5, 0.8\}} P(X = h \mid Z = z)P(Y = h \mid Z = z)p(Z = z)$$
Example: Coins (4)

- $P(Z = 0.3) = 0.7$, $P(Z = 0.5) = 0.2$ and $P(Z = 0.8) = 0.1$
- Is $P(X = Heads, Y = Heads) = P(X = Heads)p(Y = Heads)$?
  
  \begin{align*}
P(X = h, Y = h) &= \sum_{z \in \{0.3, 0.5, 0.8\}} P(X = h, Y = h | Z = z)p(Z = z) \\
&= \sum_{z \in \{0.3, 0.5, 0.8\}} P(X = h | Z = z)P(Y = h | Z = z)p(Z = z) \\
&= P(X = h | Z = 0.3)P(Y = h | Z = 0.3)p(Z = 0.3) \\
&\quad + P(X = h | Z = 0.5)P(Y = h | Z = 0.5)p(Z = 0.5) \\
&\quad + P(X = h | Z = 0.8)p(Y = h | Z = 0.8)p(Z = 0.8) \\
&= 0.3 \times 0.3 \times 0.7 + 0.5 \times 0.5 \times 0.2 + 0.8 \times 0.8 \times 0.1 \\
&= 0.177 \neq 0.39 \times 0.39 = 0.1521
  \end{align*}
Example: Coins (4)

- Let $Z$ be the bias of the coin, with $\mathcal{Z} = \{0.3, 0.5, 0.8\}$ and probabilities $P(Z = 0.3) = 0.7$, $P(Z = 0.5) = 0.2$ and $P(Z = 0.8) = 0.1$.
- Let $X$ and $Y$ be two consecutive flips of the coin.
- **Question:** Are $X$ and $Y$ conditionally independent given $Z$?
  - i.e., $P(X = x, Y = y | Z = z) = P(X = x | Z = z)P(Y = y | Z = z)$
- **Question:** Are $X$ and $Y$ independent?
  - i.e. $P(X = x, Y = y) = P(X = x)P(Y = y)$
The Distribution Changes Based on What We Know

• The coin has some true bias $z$

• If we know that bias, we reason about $P(X = x | Z = z)$
  • Namely, the probability of $x$ given we know the bias is $z$

• If we know do not know that bias, then from our perspective the coin outcomes follows probabilities $P(X = x)$
  • The world still flips the coin with bias $z$

• Conditional independence is a property of the distribution we are reasoning about, not an objective truth about outcomes
A bit more intuition

• If we know do not know that bias, then from our perspective the coin outcomes follows probabilities $P(X = x, Y = y)$
• and $X$ and $Y$ are correlated

• If we know $X = h$, do we think it’s more likely $Y = h$? i.e., is $P(X = h, Y = h) > P(X = h, Y = t)$?
My brain hurts, why do I need to know about coins?

- i.e., how is this relevant
- Let’s imagine you want to infer (or learn) the bias of the coin, from data
  - data in this case corresponds to a sequence of flips $X_1, X_2, \ldots, X_n$
- You can ask: $P(Z = z | X_1 = H, X_2 = H, X_3 = T, \ldots, X_n = H)$
More uses for independence and conditional independence

• If I told you $X = \text{roof type}$ was independent of $Y = \text{house price}$, would you use $X$ as a feature to predict $Y$?

• Imagine you want to predict $Y = \text{Has Lung Cancer}$ and you have an indirect correlation with $X = \text{Location}$ since in Location 1 more people smoke on average. If you could measure $Z = \text{Smokes}$, then $X$ and $Y$ would be conditionally independent given $Z$.

• Suggests you could look for such causal variables, that explain these correlations

• We will see the utility of conditional independence for learning models
Expected Value

The expected value of a random variable is the weighted average of that variable over its domain.

Definition: Expected value of a random variable

\[ \mathbb{E}[X] = \begin{cases} \sum_{x \in \mathcal{X}} xp(x) & \text{if } X \text{ is discrete} \\ \int_{\mathcal{X}} xp(x) \, dx & \text{if } X \text{ is continuous.} \end{cases} \]
Relationship to Population Average and Sample Average

- Or Population Mean and Sample Mean
- Population Mean = Expected Value, Sample Mean estimates this number
- e.g., Population Mean = average height of the entire population
- For RV $X = \text{height}$, $p(x)$ gives the probability that a randomly selected person has height $x$
- Sample average: you randomly sample $n$ heights from the population
  - implicitly you are sampling heights proportionally to $p$
- As $n$ gets bigger, the sample average approaches the true expected value
Expected Value with Functions

The expected value of a function $f : \mathcal{X} \to \mathbb{R}$ of a random variable is the **weighted average** of that function's value over the domain of the variable.

**Definition:** Expected value of a function of a random variable

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

**Example:**
Suppose you get $10 if heads is flipped, or lose $3 if tails is flipped. What are your winnings **on expectation**?
Expected Value Example

Example:
Suppose you get $10 if heads is flipped, or lose $3 if tails is flipped. What are your winnings on expectation?

$X$ is the outcome of the coin flip, 1 for heads and 0 for tails

$$f(x) = \begin{cases} 
3 & \text{if } X = 0 \\
10 & \text{if } X = 1 
\end{cases}$$

$Y = f(X)$ is a new random variable

$$\mathbb{E}[Y] = \mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} f(x)p(x) = f(0)p(0) + f(1)p(1) = .5 \times 3 + .5 \times 10 = 6.5$$
Expected Value is a Lossy Summary

\[ \mathbb{E}[X] = 3 \]
\[ \mathbb{E}[X^2] \approx 10 \]

\[ \mathbb{E}[X] = 3 \]
\[ \mathbb{E}[X^2] \approx 12 \]
Definition:
The expected value of $Y$ conditional on $X = x$ is

$$
\mathbb{E}[Y \mid X = x] = \begin{cases} 
\sum_{y \in \mathcal{Y}} yp(y \mid x) & \text{if } Y \text{ is discrete}, \\
\int_{\mathcal{Y}} yp(y \mid x) \, dy & \text{if } Y \text{ is continuous}.
\end{cases}
$$
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
- $\mathbb{E}[Y|X = 0]$ is different from $\mathbb{E}[Y|X = 1]$
  - e.g. $\mathbb{E}[Y|X = 0] = 70$ is different from $\mathbb{E}[Y|X = 1] = 150$
- Another example: $\mathbb{E}[X|Z = 0.3]$ the expected outcome of the coin flip given that the bias is 0.3 ($\mathbb{E}[X|Z = 0.3] = 0 \times 0.7 + 1 \times 0.3 = 0.3$)
Conditional Expectation Example (cont)

• What do we mean by $p(y \mid X = 0)$? How might it differ from $p(y \mid X = 1)$

Lots of shorter books  Lots of medium length books  A long tail, a few very long books
Conditional Expectation Example (cont)

• What do we mean by $p(y \mid X = 0)$? How might it differ from $p(y \mid X = 1)$

• $\mathbb{E}[Y \mid X = 0]$ 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 $p(y \mid X = 1)$
Conditional Expectations

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

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

**Question:** What is $\mathbb{E}[Y \mid X]$?
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 \( X, Y \):
  \( \mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \)

- Law of Total Expectation:
  \( \mathbb{E} \left[ \mathbb{E} \left[ Y \mid X \right] \right] = \mathbb{E}[Y] \)

**Question:** How would you prove these?

\[
\mathbb{E}[Y] = \sum_{y \in Y} yp(y) \\
= \sum_{y \in Y} \sum_{x \in X} p(x, y) \quad \text{def. marginal distribution}
\]

\[
= \sum_{x \in X} \sum_{y \in Y} yp(x, y) \quad \text{rearrange sums}
\]

\[
= \sum_{x \in X} \sum_{y \in Y} yp(y \mid x)p(x) \quad \text{Chain rule}
\]

\[
= \sum_{x \in X} \left( \sum_{y \in Y} yp(y \mid x) \right) p(x) \quad \text{def. } \mathbb{E}[Y \mid X = x]
\]

\[
= \sum_{x \in X} \left( \mathbb{E}[Y \mid X = x] \right) p(x) \quad \text{def. expected value of function}
\]

\[
= \mathbb{E} \left( \mathbb{E}[Y \mid X] \right) \quad \blacksquare
\]
**Variance**

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

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

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

Equivalently,

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

*(Exercise: Show that this is true)*
Covariance

**Definition:** The covariance of two random variables is

\[
\text{Cov}(X, Y) = \mathbb{E} \left[ (X - \mathbb{E}[X])^2 \right] \\
= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].
\]

**Question:** What is the range of Cov\((X, Y)\)?
Correlation

**Definition:** The correlation of two random variables is

\[
\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}
\]

**Question:** What is the range of Corr\((X, Y)\)?

**hint:** \(\text{Var}(X) = \text{Cov}(X, X)\)
Properties of Variances

- $\text{Var}[c] = 0$ 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 $X, Y$, $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$ (why?)
Independence and Decorrelation

- Recall if $X$ and $Y$ are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- Independent RVs have zero correlation (why?)
  
  hint: $\text{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

- Uncorrelated RVs (i.e., $\text{Cov}(X, Y) = 0$) might be dependent
  (i.e., $p(x, y) \neq p(x)p(y)$).
- Correlation (Pearson's correlation coefficient) shows linear relationships; but can miss nonlinear relationships

- **Example:** $X \sim \text{Uniform}\{-2, -1, 0, 1, 2\}$, $Y = X^2$
  
  - $\mathbb{E}[XY] = .2(-2 \times 4) + .2(2 \times 4) + .2(-1 \times 1) + .2(1 \times 1) + .2(0 \times 0)$
  - $\mathbb{E}[X] = 0$
  - So $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0\mathbb{E}[Y] = 0$
• **Random variables** takes different values with some probability

• The value of one variable can be informative about the value of another
  • Distributions of multiple random variables are described by the **joint** probability distribution (joint PMF or joint PDF)
  • You can have a new distribution over one variable when you **condition** on the other

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

• The **variance** of a random variable is the expected squared distance from the mean

• The **covariance** and **correlation** of two random variables can summarize how changes in one are informative about changes in the other.
Exercise applying your knowledge

• Let’s revisit the commuting example, and assume we collect continuous commute times

• We want to model commute time as a Gaussian

• What parameters do I have to specify (or learn) to model commute times with a Gaussian?

• Is a Gaussian a good choice?

\[ p(\omega) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\omega-\mu)^2} \]
Exercise applying your knowledge

- A better choice is actually what is called a Gamma distribution
Exercise applying your knowledge

- We can also consider conditional distributions \( p(y \mid x) \)
- \( Y \) is the commute time, let \( X \) be the month
- Why is it useful to know \( p(y \mid X = \text{Feb}) \) and \( p(y \mid X = \text{Sept}) \)?
- What else could we use for \( X \) and why pick it?
Exercise applying your knowledge

• Let use a simple $X$, where it is 1 if it is slippery out and 0 otherwise

• Then we could model two Gaussians, one for the two types of conditions

\[
p(y|X = 0) = \mathcal{N} \left( \mu_0, \sigma_0^2 \right)
\]

\[
p(y|X = 1) = \mathcal{N} \left( \mu_1, \sigma_1^2 \right)
\]
Exercise applying your knowledge

- Eventually we will see how to model the distribution over $Y$ using functions of other variables (features) $X$

\[
p(y|x) = \mathcal{N} \left( \mu = \sum_{j=1}^{d} w_i x_i, \sigma^2 \right)
\]