1 Conditional Expectation

Definition 1. Given an event $E$ and a random variable $X$, the expectation of $X$ conditioned on $E$ is defined as

$$E[X|E] := \sum_x P(X = x|E) \cdot x$$

In this case, it’s like we now have a new probability space which only contain the situation when $E$ happened.

Example 1. When we roll a fair dice, and we only count for the even outcome, which is $E = \{2, 4, 6\}$. In this case, each outcome has a $\frac{1}{3}$ probability to show up. That is, $P(X = 2|E) = P(X = 4|E) = P(X = 6|E) = \frac{1}{3}$.

Suppose we have another random variable $Y$ and a value $y$ that $Y$ can take, then we have

$$E[X|Y = y] = \sum_x P(X = x|Y = y) \cdot x.$$

Above, we treat $Y = y$ as an event. However, often we use shortcut notion $E[X|Y]$, which seems to be not well-defined. We should think of it as $E[X|Y = Y]$, where the first $Y$ is the random variable $Y$ and the second $Y$ is a value that the random variable can take and thus the expectation is a function of the second $Y$. So, $E[X|Y]$ is a function of $Y$.

Example 2. We have 4 balls, and each of them has their color and weight.

<table>
<thead>
<tr>
<th></th>
<th>Color</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ball 1</td>
<td>“Red”</td>
<td>3</td>
</tr>
<tr>
<td>Ball 2</td>
<td>“Red”</td>
<td>5</td>
</tr>
<tr>
<td>Ball 3</td>
<td>“Blue”</td>
<td>2</td>
</tr>
<tr>
<td>Ball 4</td>
<td>“Blue”</td>
<td>7</td>
</tr>
</tbody>
</table>

We randomly pick a ball, use $X$ to refer to the weight of ball and $Y$ for the color of ball.

$$E[X|Y = “Red”] = \frac{1}{2} \times 5 + \frac{1}{2} \times 3 = 4$$

$$E[X|Y = “Blue”] = \frac{1}{2} \times 2 + \frac{1}{2} \times 7 = 4.5$$

$$E[X|Y] = \begin{cases} 4 & \text{if } Y = “Red” \\ 4.5 & \text{if } Y = “Blue” \end{cases}$$

So, $E[X|Y]$ is a function of $Y$ that maps “Red” to 4 and “Blue” to 4.5.

2 Variance

Example 3. If we have two bank accounts both have 1000 dollars. One of them can either be added or decreased by 1 dollar each month, and the other one can be added or decreased by 500 dollars each month. We use $X_1$ and $X_2$ to represent the money in each account after one month.

$$X_1 = \begin{cases} 999 & \text{with probability } \frac{1}{2} \\ 1001 & \text{with probability } \frac{1}{2} \end{cases}$$
\[ X_2 = \begin{cases} 
500 & \text{with probability } \frac{1}{2}, \\
1500 & \text{with probability } \frac{1}{2}. 
\end{cases} \]

\( \mathbb{E}[X_1] = \mathbb{E}[X_2] = 1000. \) However \( X_2 \) is more risky, since it is more likely to deviate from the expectation.

Suppose we have a randomized variable \( X \) with \( \mu = \mathbb{E}[X] \). One way to capture the deviation of \( X \) from its expectation is to use \( \mathbb{E}[|X - \mu|] \). However, absolute value function makes the definition hard to use in many cases since it does not have a continuous derivative at \( X = \mu \). So instead we use:

**Definition 2 (Variance).** *The variance of a random variable \( X \), denoted as \( \text{Var}[X] \) is defined as*

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

In the former example, \( \text{Var}[X_1] = 1^2 = 1 \), and \( \text{Var}[X_2] = 500^2 = 2500 \). However, the squaring changed the unit. To get a quantity with the same unit as \( X \), we define the standard deviation:

**Definition 3 (Standard Deviation).** *The standard deviation of a random variable \( X \) is defined as \( \sqrt{\text{Var}[X]} \).*

Often, we use \( \sigma \) to denote the standard deviation of a random variable.

**Example 3.** We have a biased coin toss

\[
X = \begin{cases} 
1 & \text{with probability } p, \\
0 & \text{with probability } 1 - p. 
\end{cases}
\]

The expectation for \( X \) is \( \mathbb{E}[X] = p \), and the variance for \( X \) is

\[
\]

Standard deviation of \( X \) is: \( \sigma[X] = \sqrt{p(1 - p)} \).

**Lemma 4.**

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

**Proof.**

\[
\text{Var}[X] = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2 - 2\mu X + \mu^2] = \mathbb{E}[X^2] - 2\mathbb{E}[\mu X] + \mathbb{E}[\mu^2] = \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2 = \mathbb{E}[X^2] - 2\mu^2 + \mu^2 = \mathbb{E}[X^2] - \mu^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2. \quad \Box
\]

For the basic coin toss example, we have:

\[
\text{Var}[X] = \mathbb{E}[X^2] - p^2 = p \cdot 1^2 + (1 - p) \cdot 0^2 - p^2 = p(1 - p)
\]

**Definition 5.** *Definition of conditional variance:*

\[
\text{Var}[X|E] = \mathbb{E}[(X - \mathbb{E}[X|E])^2|E]
\]
3 Common Distributions

3.1 Bernoulli distribution

Bernoulli distribution is the formal name of the biased coin toss example we showed above.

**Definition 6.** Then the Bernoulli distribution with parameter \( p \in [0, 1] \) is defined as

\[
X = \begin{cases} 
1 & \text{with probability } p \\
0 & \text{with probability } 1 - p.
\end{cases}
\]

3.2 Binomial distribution

**Definition 7.** The binomial distribution \( X \) with parameters \( n \in \mathbb{Z} > 0 \) and \( p \in [0, 1] \) is the sum of \( n \) independent Bernoulli random variables with parameter \( p \). Then, for every \( i \in \{0, 1, 2, \ldots, n\} \), we have

\[
P(X = i) = \binom{n}{i} p^i (1 - p)^{n-i}.
\]

The expectation of the above random variable \( X \) is \( E[X] = np \).

**Proof.** \( X = \sum_{i=1}^{n} X_i \), where \( X_i \) is the result of the \( i \)-th coin toss. For every \( i \), \( E[X_i] = p \),

\[
E[X] = E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i] = np.
\]

A generalization of the following lemma gives that \( \text{Var}[X] = \sum_{i=1}^{n} \text{Var}[X_i] = np(1 - p) \).

**Lemma 8.** If \( X \) and \( Y \) are independent random variables, then we have \( \text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] \).

**Proof.** Suppose \( \mu_X = E[X] \), and \( \mu_Y = E[Y] \)

\[
\text{Var}[X + Y] = \text{E}[(X + Y) - \mu_X - \mu_Y]^2 = \text{E}[(X - \mu_X)^2 + (Y - \mu_Y)^2] + \text{E}[2(X - \mu_X)(Y - \mu_Y)] = \text{Var}[X] + \text{Var}[Y] + 2 \text{E}[X - \mu_X] \text{E}[Y - \mu_Y] = \text{Var}[X] + \text{Var}[Y].
\]

The first and the third equalities used the definition of variance. The second equality used the linearity of expectation.

3.3 Geometric distribution

The geometric distribution with parameter \( p \in (0, 1) \) is the distribution on the number of Bernoulli trials with parameter \( p \) needed in order to get 1 success. Then, let \( X \) be the random variable from the distribution, we have

\[
\text{Pr}[X = i] = (1 - p)^{i-1} p, \quad \forall \ i = 1, 2, 3, \ldots
\]

As we already showed, we have \( E[X] = \frac{1}{p} \). The variance of \( X \) is \( \text{Var}[X] = \frac{1-p}{p^2} \).

4 Birthday Paradox

There are 365 different possibility for a person’s birthday, and now we have \( n \) people in the classroom, each person will have a birthday that’s uniformly distributed over 365 days. What is the smallest number \( n \) such that the probability that 2 people in the room have the same birthday?

To solve this problem, we first solve a slightly different problem. Let us consider the expected number of collision pairs: A collision pair is a pair of people with the same birthday. We are
interested in how big should $n$ be in order for the expectation to be at least 1. For a pair $u \neq v$ of people, define $X_{\{u,v\}} = 1$ if $u$ and $v$ is a collision pair and 0 otherwise. Then,

$$E[\text{number of collision pairs}] = E\left[ \sum_{\{u,v\}} X_{\{u,v\}} \right] = \sum_{\{u,v\}} E[X_{\{u,v\}}] = \sum_{\{u,v\}} \frac{1}{365} = \frac{\binom{n}{2}}{365} \approx \frac{n^2}{730}.$$ 

In order for the number to be at least 1, we need $n \approx \sqrt{365 \times 2} = \sqrt{730} \approx 27$.

We can also try to compute the threshold $n$ for which the collision probability exceeds 50%. But there is also a way to approximate the threshold for the first problem. For the sake of simplicity and generalization, let us define $M = 365$. Then the probability that there is no collision is exactly

$$\frac{M}{M} \times \frac{M-1}{M} \times \frac{M-2}{M} \times \cdots \times \frac{M-n+1}{M} = 1 \times (1 - \frac{1}{M}) \times (1 - \frac{2}{M}) \times \cdots \times (1 - \frac{n-1}{M}).$$

This holds since the first person has $M$ choices to avoid a collision, the second person has $M - 1$ choices to avoid a collision, and so on. Notice that if $1 - x \approx e^{-x}$ for $x$ very close to 0.

If $n$ is much smaller than $M$, then the above quantity can be approximated by

$$e^0 \times e^{-1/M} \times e^{-2/M} \times \cdots \times e^{-(n-1)/M} = e^{-n(n-1)/(2M)}.$$

So for the probability to be at most 50%, we need $n(n-1)/(2M)$ to be $\ln 2$. As before, we also get that it suffices for $n$ to be of order $\sqrt{M}$. 
