3. Conditional Expectation and Variance

Bertsekas & Tsitsiklis: Section 4.3.

Assumed knowledge: Conditional expectation and variance, joint distributions.

Conditional Expectation

You have seen conditional expectation in Chapter 2 of the book. There, they took the form of constants. For example, $E(X)$ is a constant, $E(X|A)$ is a constant for some given event $A$, and $E(X|Y = y)$ is a constant for some given value $y$.

If you think about the last one of those examples, it makes sense to define a function $E(X|Y)$ of $Y$. Since $Y$ is a random variable, that makes $E(X|Y)$ also a random variable. One of the most common mistakes made by students in this topic is to be confused about when these expectations are constants and when they are random variables. So, pay extra attention to that.

In these notes I will use subscripts for the expectations. These are not strictly necessary in these cases, but you may find them helpful. Otherwise just ignore them. I made them really tiny. We define the conditional expectation

$$E_{X}(X|Y) = \begin{cases} 
\sum_{x} xp_{X|Y}(x|Y) & \text{discrete} \\
\int_{-\infty}^{\infty} x f_{X|Y}(x|Y) dx & \text{continuous}.
\end{cases}$$

Please go through examples 4.16-4.19 to see examples of how this works in practice. (We may do some of those in the lecture.) The following is a very useful result, because there are many situations where $E_{X}(X|Y)$ is much easier to find than $E(X)$ directly (because we know more about conditional distribution of $X$ than the marginal one).

Lemma 3.1. Law of Iterated Expectations (Law of Total Expectation):

$$E_{X}(X) = E_{Y}(E_{X}(X|Y)).$$
Proof. This is not hard to show if we remember how joint distributions work. For the discrete case, we have
\[
\mathbb{E}_Y(\mathbb{E}_X(X|Y)) = \sum_y \mathbb{E}_x(X|Y = y)p_Y(y)
\]
\[
= \sum_y \left( \sum_x x p_{X|Y}(x|y) \right) p_Y(y)
\]
\[
= \sum_x x \left( \sum_y p_{X|Y}(x|y)p_Y(y) \right)
\]
\[
= \sum_x x \left( \sum_y p_{X,Y}(x,y) \right)
\]
\[
= \sum_x x p_X(x)
\]
\[
= \mathbb{E}_x(X).
\]

For the continuous case, we have
\[
\mathbb{E}_Y(\mathbb{E}_X(X|Y)) = \int_{-\infty}^{+\infty} \mathbb{E}_x(X|Y = y)f_Y(y)dy
\]
\[
= \int_{-\infty}^{+\infty} \left( \int_{-\infty}^{+\infty} x f_{X|Y}(x|y)dx \right)f_Y(y)dy
\]
\[
= \int_{-\infty}^{+\infty} x \left( \int_{-\infty}^{+\infty} f_{X|Y}(x|y)f_Y(y)dy \right)dx
\]
\[
= \int_{-\infty}^{+\infty} x \left( \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy \right)dx
\]
\[
= \int_{-\infty}^{+\infty} x f_X(x)dx
\]
\[
= \mathbb{E}_x(X).
\]

□

Conditional Expectation as an Estimator

We define an estimator of $X$ given $Y$ as $\hat{X} = \mathbb{E}_x(X|Y)$, and the estimation error as $\tilde{X} = X - \hat{X}$.

I don’t have much to add to the book here (pages 225-226), except that:

Correction 3.2. Page 225. $\tilde{X} = \hat{X} - X$ should be $\tilde{X} = X - \hat{X}$.

I would also like to add that in many other circumstances, we use $\hat{X}$ to mean just $\mathbb{E}(X)$, so it is important to be aware of the context that this symbol is used in.

Example 3.3. Please try the following question yourself using the hints from the lecture. Let $X$ and $Y$ be continuous random variables with the following joint PDF:
\[
f_{X,Y}(x,y) = \begin{cases} 
2 & \text{if } x \geq 0, y \geq 0, y + x \leq 1, \\
0 & \text{o/w}.
\end{cases}
\]
Can you find $\mathbb{E}(X)$? Does it help to find $\mathbb{E}(X|Y)$ first? How does the Law of Iterated Expectations help?

**Example 3.4.** Now, find $\mathbb{E}(X)$ and $\mathbb{E}(Y)$ if $f_{X,Y}(x,y)$ is uniform over the triangle spanned by the points $(0,0) - (\frac{1}{2}, 0) - (0,2)$.

**Conditional Variance**

We define the **conditional variance**

$$
\text{var}_x(X|Y) = \begin{cases} 
\sum_x (x - \hat{X})^2 p_{X|Y}(x|Y) & \text{discrete} \\
\int_{-\infty}^{+\infty} (x - \hat{X})^2 f_{X|Y}(x|Y) dx & \text{continuous}.
\end{cases}
$$

A few facts about estimators and errors are worth highlighting, before we move on to the main result. Besides checking these, try to explain intuitively why they make sense, if $\hat{X}$ is an estimator of $X$ based on knowledge of the random variable $Y$.

**Example 3.5.** Prove that $\mathbb{E}(\hat{X}) = 0$.

**Example 3.6.** Prove that $\text{cov}(\hat{X}, \hat{X}) = 0$.

**Example 3.7.** Prove that $\text{cov}(\hat{X}, Y) = 0$.

The following is a powerful tool that helps us decompose the variance of a random variable $X$ in terms of variance explained by another random variable $Y$, and variance not explained by $Y$. As in the example of the students and their scores (Ex. 4.18): if we look at the variance between student grades, then this variance is either explained by variance between the different sections, or by variance within each the sections. Of course in most cases, it is a combination of both.

**Lemma 3.8.** Law of Total Variance (Variance Decomposition Formula):

$$
\text{var}(X) = \mathbb{E}_y(\text{var}_x(X|Y)) + \text{var}_y(\mathbb{E}_x(X|Y)).
$$

**Proof.** This is a dense version of the proof in the book. You may wish to practice by filling in the missing parts. This result follows from

- $\text{var}(X) = \text{var}_x(\hat{X} + \hat{X}) = \text{var}_x(\hat{X}) + \text{var}_x(\hat{X})$
  
  Proof:
  
  $\mathbb{E}_x(\hat{X}) = \mathbb{E}_y(\mathbb{E}_x(\hat{X}|Y)) = \mathbb{E}_y(\mathbb{E}_x(X - \hat{X}|Y)) = \mathbb{E}_y(\mathbb{E}_x(X|Y) - \mathbb{E}_x(\hat{X}|Y)) = \mathbb{E}_y(\hat{X} - \hat{X}) = 0$
  
  $\mathbb{E}_x(\hat{X}\hat{X}) = \ldots = 0$ (exercise)
  
  $\Rightarrow\text{cov}(\hat{X}, \hat{X}) = 0$.

- $\text{var}_x(\hat{X}) = \mathbb{E}_y(\text{var}_x(X|Y))$
  
  Proof:
  
  $\text{var}_x(X|Y) = \mathbb{E}_x(\hat{X}^2|Y)$
Here is an alternative way of writing the proof of Lemma 3.8, which we used in the lecture.

**Proof.** We write
\[ X - \mathbb{E}(X) = (X - \hat{X}) + (\hat{X} - \mathbb{E}(X)) \]
(square and take expectation on both sides)
\[ \text{var}(X) = \mathbb{E}\left( (X - \hat{X})^2 \right) + \mathbb{E}\left( (\hat{X} - \mathbb{E}(X))^2 \right) + 2\mathbb{E}\left( (X - \hat{X}) \cdot (\hat{X} - \mathbb{E}(X)) \right). \]

First, by total expectation,
\[ \mathbb{E}\left( (X - \hat{X})^2 \right) = \mathbb{E}\left( \mathbb{E}\left( (X - \hat{X})^2 \right | Y \right) = \mathbb{E}(\text{var}(X | Y)). \]

Secondly, because \( \mathbb{E}(\hat{X}) = \mathbb{E}(X) \),
\[ \mathbb{E}\left( (\hat{X} - \mathbb{E}(X))^2 \right) = \text{var}(\mathbb{E}(X | Y)). \]

Finally, we let \( h(Y) = 2(\hat{X} - \mathbb{E}(X)) \). It is easy to check that \( h(Y) \) is a function of \( Y \) only (constant in terms of \( X \)). We have
\[ 2\mathbb{E}\left( (X - \hat{X}) \cdot (\hat{X} - \mathbb{E}(X)) \right) = \mathbb{E}\left( (X - \hat{X})h(Y) \right) \]
\[ = \mathbb{E}(Xh(Y)) - \mathbb{E}(\hat{X})h(Y) \]
\[ = \mathbb{E}(Xh(Y)) - \mathbb{E}(\mathbb{E}(X | Y)h(Y)) \]
\[ = \mathbb{E}(Xh(Y)) - \mathbb{E}(Xh(Y)) \]
\[ = 0. \]

This completes the proof.

\[ \square \]

**Code for Sage/Wolfram Alpha**

Wolfram Alpha does not allow you to directly input distribution functions, but it does have many built in distributions, and allows you to condition on certain events. For example:

- probability of 8 successes in 14 trials with p=.6
- probability that \( x^2 > 2 \) given that \( x > 1 \), \( x \) standard normal
- \( \mathbb{E}[x^2] \) where \( x \) is exponentially distributed

**Solutions to Exercises**

**Solution 3.5** We want to check that \( \mathbb{E}(\tilde{X}) = 0 \). We have
\[ \mathbb{E}(\tilde{X}) = \mathbb{E}(X - \hat{X}) = \mathbb{E}(X) - \mathbb{E}(\hat{X}) = \mathbb{E}(X) - \mathbb{E}(X) = 0. \]
Solution 3.6 We want to check that that $\text{cov}(\tilde{X}, \hat{X}) = 0$. We have

$$\text{cov}(\tilde{X}, \hat{X}) = \text{cov}(\tilde{X}, \hat{X}) = \mathbb{E}(\tilde{X}\hat{X}) - \mathbb{E}(\tilde{X})\mathbb{E}(\hat{X}) = \mathbb{E}(\tilde{X}\hat{X}) - 0 = \mathbb{E}(\mathbb{E}(\tilde{X}\hat{X} | Y)) = \mathbb{E}(\tilde{X}\mathbb{E}(\hat{X} | Y)) = 0.$$ 

The last step is allowed because $\hat{X}$ is a function of $Y$ only. So, if $Y$ is known, it is a constant. We also have that

$$\mathbb{E}(\hat{X} | Y) = \mathbb{E}(X - \hat{X} | Y) = \mathbb{E}(X | Y) - \mathbb{E}(\hat{X} | Y) = 0.$$ 

Solution 3.7 Finally, we want to check that that $\text{cov}(\tilde{X}, Y) = 0$. We write

$$\text{cov}(\tilde{X}, Y) = \text{cov}(X, Y) - \text{cov}(\hat{X}, Y).$$ 

Now, we use the fact that $\mathbb{E}(\hat{X}) = \mathbb{E}(X)$, to show that

$$\text{cov}(\tilde{X}, Y) = \mathbb{E}[(\hat{X} - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))]$$

$$= \mathbb{E}(\hat{X}Y) - \mathbb{E}(\mathbb{E}(X)Y) - \mathbb{E}(\hat{X}\mathbb{E}(Y)) + \mathbb{E}(\mathbb{E}(X)\mathbb{E}(Y))$$

$$= \mathbb{E}(\mathbb{E}(XY | Y)) - \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y)$$

$$= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \text{cov}(X, Y).$$

Hence,

$$\text{cov}(\tilde{X}, Y) = 0.$$