MATH 170B LECTURE NOTE 3: CONDITIONAL EXPECTATION AND VARIANCE

HANBAEK LYU

3. How can we handle multiple source of randomness?

3.1. **Conditional expectation.** Let X, Y be discrete RVs. Recall that the expectation $\mathbb{E}(X)$ is the 'best guess' on the value of X when we do not have any prior knowledge on X. But suppose we have observed that some possibly related RV Y takes value y. What should be our best guess on X, leveraging this added information? This is called the *conditional expectation of X given $Y = y$*, which is defined by

$$\mathbb{E}[X|Y = y] = \sum_x x\,\mathbb{P}(X = x|Y = y). \tag{1}$$

This best guess on X given $Y = y$, of course, depends on y. So it is a function in y. Now if we do not know what value Y might take, then we omit y and $\mathbb{E}[X|Y]$ becomes a RV, which is called the *conditional expectation of X given Y*.

Example 3.1. Suppose we have a biased coin whose probability of heads is itself random and is distributed as $Y \sim \text{Uniform}([0,1])$. Let's flip this coin n times and let X be the total number of heads. Given that $Y = y \in [0,1]$, we know that X follows Binomial(n, y) (in this case we write $X|U \sim$ Binomial(n, Y)). So $\mathbb{E}[X|Y = y] = ny$. Hence as a random variable, $\mathbb{E}[X|Y] = nY \sim \text{Uniform}([0,n])$. So the expectation of $\mathbb{E}[X|Y]$ is the mean of Uniform$([0,n])$, which is $n/2$. This value should be the true expectation of X.

The above example suggests that if we first compute the conditional expectation of X given $Y = y$, and then average this value over all choice of y, then we should get the actual expectation of X. Justification of this observation is based on the following fact

$$\mathbb{P}(Y = y\,|\,X = x)\mathbb{P}(X = x) = \mathbb{P}(X = x,\, Y = y) = \mathbb{P}(X = x\,|\,Y = y)\mathbb{P}(Y = y). \tag{2}$$

That is, if we are interested in the event that $(X, Y) = (x, y)$, then we can either first observe the value of X and then Y, or the other way around.

Proposition 3.2 (Iterated expectation)**.** *Let X, Y be discrete RVs. Then* $\mathbb{E}(X) = \mathbb{E}[\mathbb{E}[X|Y]]$.

Proof. We are going to write the iterated expectation $\mathbb{E}[\mathbb{E}[X|Y]]$ as a double sum and swap the order of summation (Fubini's theorem, as always).

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[X|Y]] &= \sum_y \mathbb{E}[X|Y = y]\mathbb{P}(Y = y) && (3)\\
&= \sum_y \left(\sum_x x\,\mathbb{P}(X = x|Y = y)\right)\mathbb{P}(Y = y) && (4)\\
&= \sum_y \sum_x x\,\mathbb{P}(X = x|Y = y)\mathbb{P}(Y = y) && (5)\\
&= \sum_y \sum_x x\,\mathbb{P}(X = x,\, Y = y) && (6)\\
&= \sum_x \sum_y x\,\mathbb{P}(Y = y|X = x)\mathbb{P}(X = x) && (7)\\
&= \sum_x x\left(\sum_y \mathbb{P}(Y = y|X = x)\right)\mathbb{P}(X = x) && (8)\\
&= \sum_x x\,\mathbb{P}(X = x) = \mathbb{E}(X). && (9)
\end{aligned}$$

□

Remark 3.3. Here is an intuitive reason why the iterated expectation works. Suppose you want to make the best guess $\mathbb{E}(X)$. Pretending you know Y, you can imporove your guess to be $E(X\,|\,Y)$. Then you admit that you didn't know anything about Y and average over all values of Y. The result is $\mathbb{E}[\mathbb{E}[X\,|\,Y]]$, and this should be the same best guess on X when we don't know anything about Y.

All our discussions above hold for continuous RVs as well: We simply replace the sum by integral and PMF by PDF. To summarize how we compute the iterated expectations when we condition on discrete and continuous RV:

$$\mathbb{E}[\mathbb{E}[X\,|\,Y]] = \begin{cases} \sum_y \mathbb{E}[X\,|\,Y=y]\mathbb{P}(Y=y) & \text{if } Y \text{ is discrete} \\ \int_{-\infty}^{\infty} \mathbb{E}[X\,|\,Y=y] f_Y(y)\,dy & \text{if } Y \text{ is continuous.} \end{cases} \tag{10}$$

Exercise 3.4 (Iterated expectation for probability)**.** Let X, Y be RVs.

(i) For any $x \in \mathbb{R}$, show that $\mathbb{P}(X \le x) = \mathbb{E}[\mathbf{1}(X \le x)]$.

(ii) By using iterated expectation, show that

$$\mathbb{P}(X \le x) = \mathbb{E}[\mathbb{P}(X \le x\,|\,Y)], \tag{11}$$

where the expectation is taken over for all possible values of Y.

Example 3.5 (Example 3.1 revisited)**.** Let $Y \sim \text{Uniform}([0,1])$ and $X \sim \text{Binomial}(n, Y)$. Then $X|Y = y \sim \text{Binomial}(n, y)$ so $\mathbb{E}[X|Y=y] = ny$. Hence

$$\mathbb{E}[X] = \int_0^1 \mathbb{E}[X\,|\,Y=y] f_Y(y)\,dy = \int_0^1 ny\,dy = n/2. \tag{12}$$

Example 3.6. Let $X_1 \sim \text{Exp}(\lambda_1)$ and $X_2 \sim \text{Exp}(\lambda_2)$ be independent exponential RVs. We will show that

$$\mathbb{P}(X_1 < X_2) = \frac{\lambda_2}{\lambda_1 + \lambda_2} \tag{13}$$

using the iterated expectation. Using iterated expectation for probability,

$$\mathbb{P}(X_1 < X_2) = \int_0^\infty \mathbb{P}(X_1 < X_2\,|\,X_1 = x_1)\lambda_1 e^{-\lambda_1 x_1}\,dx_1 \tag{14}$$

$$= \int_0^\infty \mathbb{P}(X_2 > x_1)\lambda_1 e^{-\lambda_1 x_1}\,dx_1 \tag{15}$$

$$= \lambda_1 \int_0^\infty e^{-\lambda_2 x_1} e^{-\lambda_1 x_1}\,dx_1 \tag{16}$$

$$= \lambda_1 \int_0^\infty e^{-(\lambda_1+\lambda_2)x_1}\,dx_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2}. \tag{17}$$

Exercise 3.7. Consider a post office with two clerks. Three people, A, B, and C, enter simultaneously. A and B go directly to the clerks, and C waits until either A or B leaves before he begins service. Let X_A be the time that A spends at a register, and define X_B and X_B similarly. Compute the probability $\mathbb{P}(X_A > X_B + X_C)$ that A leaves the post office after B and C, in the following scenarios:

(a) the service time for each clerk is exactly (nonrandom) ten minutes?

(b) the service times are i, independently with probability $1/3$ for $i \in \{1,2,3\}$?

(c) the service times are independent exponential variables with mean $1/\mu$?

Exercise 3.8. Suppose we have a stick of length L. Break it into two pieces at a uniformly chosen point and let X_1 be the length of the longer piece. Break this longer piece into two pieces at a uniformly chosen point and let X_2 be the length of the longer one. Define $X_3, X_4, \cdots$ in a similar way.

(i) Show that $X_1 \sim \text{Uniform}([L/2, L])$.

(ii) Show that $X_2 \,|\, X_1 \sim \text{Uniform}([X_1/2, X_1])$.
(iii) Show that $X_{n+1} \,|\, X_n \sim \text{Uniform}([X_n/2, X_n])$.
(iv) Show that $\mathbb{E}[X_n] = (3L/4)^n$.

3.2. **Conditional expectation as an estimator.** We introduced the conditional expectation $\mathbb{E}[X \,|\, Y = y]$ as the best guess on X given that $Y = y$. Such a 'guess' on a RV is called an *estimator*. Let's first take a look at two extremal cases, where observing Y gives absolutely no information on X or gives everything.

Example 3.9. Let X and Y be independent discrete RVs. Then knowing the value of Y should not yield any information on X. In other words, given that $Y = y$, the best guess of X should still be $\mathbb{E}(X)$. Indeed,

$$\mathbb{E}(X|Y = y) = \sum_{x=0}^{n} x\mathbb{P}(X = x \,|\, Y = y) = \sum_{x=0}^{n} x\mathbb{P}(X = x) = \mathbb{E}(X). \tag{18}$$

On the other hand, given that $X = x$, the best guess on X is just x, since the RV X has been revealed and there is no further randomness. In other words,

$$\mathbb{E}(X|X = x) = \sum_{z=0}^{n} z\mathbb{P}(X = z \,|\, X = x) = \sum_{x=0}^{n} x\mathbf{1}(z = x) = x. \tag{19}$$

Exercise 3.10. Let X, Y be discrete RVs. Show that for any function $g : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[Xg(Y) \,|\, Y] = g(Y)\mathbb{E}[X \,|\, Y]. \tag{20}$$

We now observe some general properties of the conditional expectation as an estimator.

Exercise 3.11. Let X, Y be RVs and denote $\hat{X} = \mathbb{E}[X|Y]$, meaning that $\hat{X}$ is an estimator of X given Y. Let $\tilde{X} = \hat{X} - X$ be the *estimation error*.
(i) Show that $\hat{X}$ is an *unbiased* estimator of X, that is, $\mathbb{E}(\hat{X}) = \mathbb{E}(X)$.
(ii) Show that $\mathbb{E}[\hat{X}|Y] = \hat{X}$. Hence knowing Y does not improve our current best guess $\hat{X}$.
(iii) Show that $\mathbb{E}[\tilde{X}] = 0$.
(iv) Show that $\text{Cov}(\hat{X}, \tilde{X}) = 0$. Conclude that

$$\text{Var}(X) = \text{Var}(\hat{X}) + \text{Var}(\tilde{X}). \tag{21}$$

3.3. **Conditional variance.** As we have defined conditional expectation, we could define the variance of a RV X given that anther RV Y takes a particular value. Recall that the (unconditioned) variance of X is defined by

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]. \tag{22}$$

Note that there are two places where we take expectation. Given Y, we should improve both expectations so the *conditional variance of X given Y is defined by*

$$\text{Var}(X \,|\, Y) = \mathbb{E}[(X - \mathbb{E}[X \,|\, Y])^2 \,|\, Y]. \tag{23}$$

Proposition 3.12. *Let X and Y be RVs. Then*

$$\text{Var}(X \,|\, Y) = \mathbb{E}[X^2 \,|\, Y] - \mathbb{E}[X \,|\, Y]^2. \tag{24}$$

Proof. Using linearity of conditional expectation and the fact that $\mathbb{E}[X \,|\, Y]$ is not random given Y,

$$\text{Var}(X \,|\, Y) = \mathbb{E}[X^2 - 2X\mathbb{E}[X \,|\, Y] + \mathbb{E}[X \,|\, Y]^2 \,|\, Y] \tag{25}$$

$$= \mathbb{E}[X^2 \,|\, Y] - \mathbb{E}[2X\mathbb{E}[X \,|\, Y] \,|\, Y] + \mathbb{E}[\mathbb{E}[X \,|\, Y]^2 \,|\, Y] \tag{26}$$

$$= \mathbb{E}[X^2 \,|\, Y] - \mathbb{E}[X \,|\, Y]\mathbb{E}[2X \,|\, Y] + \mathbb{E}[X \,|\, Y]^2\mathbb{E}[1 \,|\, Y] \tag{27}$$

$$= \mathbb{E}[X^2 \,|\, Y] - 2\mathbb{E}[X \,|\, Y]^2 + \mathbb{E}[X \,|\, Y]^2 \tag{28}$$

$$= \mathbb{E}[X^2 \,|\, Y] - \mathbb{E}[X \,|\, Y]^2. \tag{29}$$

□

The following exercise explains in what sense the conditional expectation $\mathbb{E}[X \,|\, Y]$ is the best guess on X given Y, and that the minimum possible mean squared error is exactly the conditional variance $\mathrm{Var}(X \,|\, Y)$.

Exercise 3.13. Let X, Y be RVs. For any function $g : \mathbb{R} \to \mathbb{R}$, consider $g(Y)$ as an estimator of X. Let $\mathbb{E}_Y[(X - g(Y))^2 \,|\, Y]$ be the *mean squared error.*

(i) Show that

$$\mathbb{E}_Y[(X - g(Y))^2 \,|\, Y] = \mathbb{E}_Y[X^2 \,|\, Y] - 2g(Y)\mathbb{E}_Y[X \,|\, Y] + g(Y)^2 \tag{30}$$

$$= (g(Y) - \mathbb{E}_Y(X \,|\, Y))^2 + \mathbb{E}_Y[X^2 \,|\, Y] - \mathbb{E}_Y[X \,|\, Y]^2 \tag{31}$$

$$= (g(Y) - \mathbb{E}_Y(X \,|\, Y))^2 + \mathrm{Var}(X \,|\, Y). \tag{32}$$

(ii) Conclude that the mean squared error is minimized when $g(Y) = \mathbb{E}_Y[X \,|\, Y]$ and the global minimum is $\mathrm{Var}(X \,|\, Y)$.

Next, we study how we can decompose the variance of X by conditioning on Y.

Proposition 3.14 (Law of total variance)**.** *Let X and Y be RVs. Then*

$$\mathrm{Var}(X) = \mathbb{E}(\mathrm{Var}(X \,|\, Y)) + \mathrm{Var}(\mathbb{E}[X \,|\, Y]). \tag{33}$$

Proof. Using previous result, iterated expectation, and linearity of expectation, we have

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \tag{34}$$

$$= \mathbb{E}_Y(\mathbb{E}(X^2|Y)) - (\mathbb{E}_Y(\mathbb{E}(X|Y)))^2 \tag{35}$$

$$= \mathbb{E}_Y(\mathrm{Var}(X|Y) + (\mathbb{E}(X|Y))^2) - (\mathbb{E}_Y(\mathbb{E}(X|Y)))^2 \tag{36}$$

$$= \mathbb{E}_Y(\mathrm{Var}(X|Y)) + \left[\mathbb{E}_Y(\mathbb{E}(X|Y))^2) - (\mathbb{E}_Y(\mathbb{E}(X|Y)))^2\right] \tag{37}$$

$$= \mathbb{E}_Y(\mathrm{Var}(X|Y)) + \mathrm{Var}_Y(\mathbb{E}(X|Y)). \tag{38}$$

□

Here is a handwavy explanation on why the above is true. Given Y, we should measure the fluctuation of $X \,|\, Y$ from the conditional expectation $\mathbb{E}[X \,|\, Y]$, and this is measured as $\mathrm{Var}(X \,|\, Y)$. Since we don't know Y, we average over all Y, giving $\mathbb{E}(\mathrm{Var}(X \,|\, Y))$. But the reference point $\mathbb{E}[X \,|\, Y]$ itself varies with Y, so we should also measure its own fluctuation by $\mathrm{Var}(\mathbb{E}[X \,|\, Y])$. These fluctuations add up nicely like Pythagorian theorem because $\mathbb{E}[X \,|\, Y]$ is an optimal estimator so that these two fluctuations are 'orthogonal'.

Exercise 3.15. Let X, Y be RVs. Write $\bar{X} = \mathbb{E}[X \,|\, Y]$ and $\tilde{X} = X - \mathbb{E}[X \,|\, Y]$ so that $X = \bar{X} + \tilde{X}$. Here $\bar{X}$ is the estimate of X given Y and $\tilde{X}$ is the estimation error.

(i) Using Exercise 3.11 (iii) and iterated expectation, show that

$$\mathbb{E}[\tilde{X}^2] = \mathrm{Var}(\mathbb{E}[X \,|\, Y]). \tag{39}$$

(ii) Using Exercise 3.11 (iv), conclude that

$$\mathrm{Var}(X) = \mathbb{E}(\mathrm{Var}(X \,|\, Y)) + \mathrm{Var}(\mathbb{E}[X \,|\, Y]). \tag{40}$$

Example 3.16. Let $Y \sim \mathrm{Uniform}([0,1])$ and $X \sim \mathrm{Binomial}(n, Y)$. Since $X \,|\, Y = y \sim \mathrm{Binomial}(n, y)$, we have $\mathbb{E}[X \,|\, Y = y] = ny$ and $\mathrm{Var}(X \,|\, Y = y) = ny(1-y)$. Also, since $Y \sim \mathrm{Uniform}([0,1])$, we have

$$\mathrm{Var}(\mathbb{E}[X \,|\, Y]) = \mathrm{Var}(nY) = \frac{n^2}{12}. \tag{41}$$

So by iterated expectation, we get

$$\mathbb{E}(X) = \mathbb{E}_Y(\mathbb{E}[X \mid Y]) = \int_0^1 ny\,dy = \frac{n}{2}. \tag{42}$$

On the other hand, by law of total variance,

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}(\mathrm{Var}(X \mid Y)) + \mathrm{Var}(\mathbb{E}(X \mid Y)) &(43)\\ &= \int_0^1 ny(1-y)\,dy + \mathrm{Var}(nY) &(44)\\ &= n\left[\frac{y^2}{2} - \frac{y^3}{3}\right]_0^1 + \frac{n^2}{12} &(45)\\ &= \frac{n^2}{12} + \frac{n}{6}. &(46)\end{aligned}$$

In fact, we can figure out the entire distribution of the binomial variable with uniform rate using conditioning, not just its mean and variance (credit to our TA Daniel).

Exercise 3.17. Let $Y \sim \mathrm{Uniform}([0,1])$ and $X \sim \mathrm{Binomial}(n, Y)$ as in Exercise 3.16.

(i) Use iterated expectation for probability to write

$$\mathbb{P}(X = k) = \binom{n}{k} \int_0^1 y^k (1-y)^{n-k}\,dy. \tag{47}$$

(ii) Write $A_{n,k} = \int_0^1 y^k (1-y)^y\,dy$. Use integration by parts and show that

$$A_{n,k} = \frac{k}{n-k+1} A_{n,k-1}. \tag{48}$$

for all $1 \le k \le n$. Conclude that for all $0 \le k \le n$,

$$A_{n,k} = \frac{1}{\binom{n}{k}} \frac{1}{n+1}. \tag{49}$$

(iii) Conclude that $X \sim \mathrm{Uniform}(\{0, 1, \cdots, n\})$.

Exercise 3.18 (Exercise 3.8 continued)**.** Let $X_1, X_2, \cdots, X_n$ be as in Exercise 3.8.

(i) Show that $\mathrm{Var}(X_1) = L^2/48$.
(ii) Show that $\mathrm{Var}(X_2) = (7/12)\mathrm{Var}(X_1) + (1/48)\mathbb{E}(X_1)^2$.
(iii) Show that $\mathrm{Var}(X_{n+1}) = (7/12)\mathrm{Var}(X_n) + (1/48)\mathbb{E}(X_n)^2$ for any $n \ge 1$.
(iv) Using Exercise 3.8, show the following recursion on variance holds:

$$\mathrm{Var}(X_{n+1}) = \frac{7}{12}\mathrm{Var}(X_n) + \frac{1}{48}\left(\frac{9}{16}\right)^n L^2. \tag{50}$$

Furthermore, compute $\mathrm{Var}(X_2)$ and $\mathrm{Var}(X_3)$.
(v)* Let $A_n = \left(\frac{16}{9}\right)^n \mathrm{Var}(X_n)$. Show that A_n's satisfy

$$A_{n+1} + L^2 = \left(\frac{28}{27}\right)(A_n + L^2). \tag{51}$$

(vi)* Show that $A_n = \left(\frac{28}{27}\right)^{n-1}(A_1 + L^2) - L^2$ for all $n \ge 1$.
(vii)* Conclude that

$$\mathrm{Var}(X_n) = \left[\left(\frac{7}{12}\right)^n - \left(\frac{9}{16}\right)^n\right] L^2. \tag{52}$$