1 Conditional Expectation

We saw that conditional probabilities represent how our beliefs for probabilities of events change given more information. The conditional expectation is how the belief of our expected value of a random variable changes given more information, often in the form of information from a related random quantity.

1.1 Conditional Distributions

Recall that for events A, B with $\mathbb{P}(B) \neq 0$ we defined

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

This gives the following natural definition for random variables.

Definition 1 (Conditional Probability Mass Function)**.** The *conditional PMF* of X given $Y = y$ is

$$p_{X \mid Y}(x \mid y) = \mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{p_{X,Y}(x,y)}{p_Y(y)} \text{ provided that } p_Y(y) > 0.$$

Similarly, the *conditional probability mass function* of Y given $X = x$ is

$$p_{Y \mid X}(y \mid x) = \mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)} = \frac{p_{X,Y}(x,y)}{p_X(x)}, \text{ provided that } p_X(x) > 0.$$

For each fixed y, the function $p_{X \mid Y}(x \mid y)$ is the probability mass function of the random variable $X \mid Y = y$ and has the usual properties, such as summing to 1. We can define the conditional PDF in the analogous way even though PDFs are not necessarily probabilities.

Definition 2 (Conditional Probability Density Function)**.** The *conditional PDF* of X given $Y = y$ is

$$f_{X \mid Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \text{ provided that } f_Y(y) > 0.$$

Similarly, the *conditional PDF* of Y given $X = x$ is

$$f_{Y \mid X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \text{ provided that } f_X(x) > 0.$$

From the conditional probability functions, we get the analogues of the Bayes Rule and the law of total probability.

Theorem 1 (*Bayes' Rule*)

1. *For discrete random variables X and Y*

$$p_{Y \mid X}(y \mid x) = \frac{p_{X \mid Y}(x \mid y) p_Y(y)}{p_X(x)}. \qquad \mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(X = x \mid Y = y)\, \mathbb{P}(Y = y)}{\mathbb{P}(X = x)}$$

2. *For continuous random variables X and Y*

$$f_{Y \mid X}(y \mid x) = \frac{f_{X \mid Y}(x \mid y) f_Y(y)}{f_X(x)}$$

Theorem 2 (*Law of Total Probability*)

1. *For discrete random variables X and Y*

$$p_X(x) = \sum_{y \in \mathcal{Y}} p_{X \mid Y}(x \mid y) p_Y(y). \qquad \mathbb{P}(X = x) = \sum_{y \in Y(\Omega)} \mathbb{P}(X = x \mid Y = y)\, \mathbb{P}(Y = y).$$

2. *For continuous random variables X and Y*

$$f_X(x) = \int_{-\infty}^{\infty} f_{X \mid Y}(x \mid y) f_Y(y)\, dy.$$

1.2 Conditional Expectation w.r.t. Random Variables

Throughout this section, we assume that X given $Y = y$ is a continuous random variable with density function $f_{X|Y}(\cdot|y)$ (if $X|Y$ is discrete, replace all the integral signs by summation signs). The conditional expectation of X given $Y = y$ is given by the expected value with respect to the conditional density function

$$\mathbb{E}[X | Y = y] = \int_{\mathbb{R}} x f_{X|Y}(x|y)\, dx.$$

This motivates the following definition:

Definition 3 (Conditional Expectation)**.** The *conditional expectation* of X given Y is the random variable

$$\mathbb{E}[X|Y] = \int_{\mathbb{R}} x f_{X|Y}(x|Y)\, dx.$$

Remark 1. The conditional expectation is a random variable since it takes elements in the domain of Y and assigns it to a number. In other words, if we define the function g through

$$g(y) = \mathbb{E}[X | Y = y] = \int_{\mathbb{R}} x f_{X|Y}(x|y)\, dx,$$

then, with a slight abuse of notation, we define

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

We can interpret the conditional expected value as the "best" estimate for the value of X given a realization of Y (see Problem 1.12). This also leads to a nice geometric interpretation of the conditional expectation (see Problem 1.2)

The conditional expectation can be used to compute complicated expected values by breaking it into more manageable pieces. This is because of the following properties.

Proposition 1

The conditional expectation has the following properties:

1. *Independence: If X and Y are independent, then*

$$\mathbb{E}[X \mid Y] = \mathbb{E}[X] \quad \text{and} \quad \mathbb{E}[Y \mid X] = \mathbb{E}[Y].$$

2. *Linearity: For any random variables X_1 and X_2,*

$$\mathbb{E}[X_1 + X_2 \mid Y] = \mathbb{E}[X_1 \mid Y] + \mathbb{E}[X_2 \mid Y]$$

3. *Pulling out known factors: If h is a function, then*

$$\mathbb{E}\left[h(Y)X|Y\right] = h(Y)\mathbb{E}\left[X|Y\right]$$

4. *Law of total expectation: $\mathbb{E}\left[\mathbb{E}\left[X|Y\right]\right] = \mathbb{E}\left[X\right]$*

Remark 2. Recall that we can only pull constants $c \in \mathbb{R}$ out of the expected values in general $\mathbb{E}[cX] = c\,\mathbb{E}[X]$. The third property allows us to pull random variables out if we can condition on it.

Likewise, one can define the conditional variance in the obvious way.

Definition 4 (Conditional Variance)**.** The *conditional variance* of X given Y is defined as

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

The conditional variance can also be used to compute complicated variances values by breaking it into more manageable pieces. This is because of the following useful properties.

Proposition 2

We have

1. $\operatorname{Var}(X|Y) = \mathbb{E}\left[X^2 \mid Y\right] - \left(\mathbb{E}\left[X \mid Y\right]\right)^2$
2. *Law of total variance:* $\operatorname{Var}(X) = \mathbb{E}\left[\operatorname{Var}(X|Y)\right] + \operatorname{Var}\left(\mathbb{E}\left[X|Y\right]\right)$

1.3 Abstract Definition of Conditional Expectation

One downside with our previous motivation is that for continuous random variables, conditioning on the event $\{Y = y\}$ is a probability zero event. In many cases, conditioning on $\{Y = y\}$ gives us degenerate forms of the conditional density (see Problem 1.9). We now introduce a notion of conditional expected values that bypasses this technical issue.

Definition 5. The *conditional expected value* $\mathbb{E}[X \mid Y]$ is a function of Y such that for any Borel set $A \in \mathcal{B}$, we have

$$\mathbb{E}[\mathbb{E}[X \mid Y]\mathbb{1}(Y \in A)] = \mathbb{E}[X\mathbb{1}(Y \in A)].$$

We can prove that the function $g(Y) = \mathbb{E}[X \mid Y]$ that satisfies this consistency equation is unique (Problem 1.14). The proof of existence is beyond the scope of the course, and requires the Radon-Nikodym theorem. Furthermore, it is easy to check that this abstract definition is equivalent to Definition 3 when the joint densities exist. This gives us a very natural definition of the conditional probability in terms of conditional expected values, by taking the expected value of the indicator function.

Definition 6. The *conditional probability* is the probability measure such that for any (measurable) set $A \subset \mathbb{R}$,

$$\mathbb{P}(X \in A \mid Y) = \mathbb{E}[\mathbb{1}(X \in A) \mid Y].$$

Although it might seem like an unintuitive definition, this definition is entirely consistent with how one would define our best guess of X given Y (see Problem 1.9). Another advantage is that from these definitions, one can derive all the key properties in Proposition 1 without defining the conditional densities, so it generalizes to random variables that do not have a well defined conditional density. It also is consistent with an even more abstract definition of the conditional expected value when you condition on σ-algebras, which you will see in future courses.

1.4 Example Problems

Problem 1.1. Show that the converse of the independence property in Proposition 1 is false. That is, if $\mathbb{E}[X \mid Y] = \mathbb{E}[X]$, then X and Y may not necessarily be independent.

Solution 1.1. Let $X \sim \mathrm{N}(0, 1)$ and let $Y = X^2$. Computing the conditional distribution carefully is possible (see Remark 3), but the intuition is clear. We have that $\mathbb{E}[X \mid Y] = \mathbb{E}[X] = 0$, since knowing X^2 tells you the magnitude of X, but not its sign. If we know that $Y = X^2 = y$, then X can only take two values $\pm\sqrt{y}$, with probability 0.5 by symmetry. We don't know which one it takes, so our best guess is the average of these two numbers, which is 0.

This is a counterexample, because although $\mathbb{E}[X \mid Y] = \mathbb{E}[X] = 0$, Y and X are not independent, since knowing X completely determines Y.

Remark 3. It is possible to show that

$$X \mid Y = y \sim \begin{cases} +\sqrt{y}, & \text{with probability } \frac{1}{2}, \\ -\sqrt{y}, & \text{with probability } \frac{1}{2}. \end{cases}$$

However, we will have to use the abstract definition of the conditional expectation because the joint distribution is neither discrete nor continuous (see Problem 1.9).

Problem 1.2. For any function h, show that

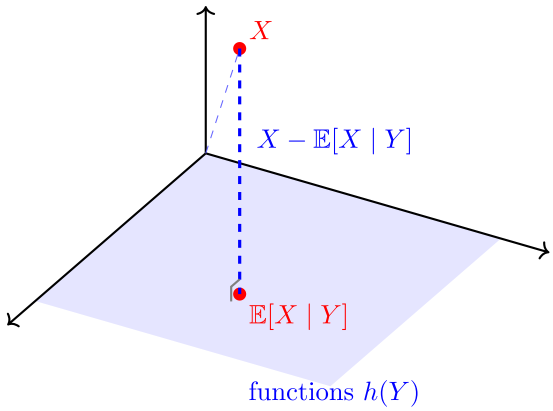
$$\mathbb{E}[(X - \mathbb{E}[X \mid Y])h(Y)] = 0.$$

That is, the random variable $Z = X - \mathbb{E}[X \mid Y]$ is uncorrelated with $h(Y)$.

Solution 1.2. We can explicitly compute this using the properties of conditional expectation,

$$\begin{aligned} \mathbb{E}[(X - \mathbb{E}[X \mid Y])h(Y)] &= \mathbb{E}[Xh(Y)] - \mathbb{E}[\mathbb{E}[X \mid Y]h(Y)] && \text{linearity} \\ &= \mathbb{E}[Xh(Y)] - \mathbb{E}[\mathbb{E}[Xh(Y) \mid Y]] && \text{pulling out known factors} \\ &= \mathbb{E}[Xh(Y)] - \mathbb{E}[Xh(Y)] = 0. && \text{law of total expectation} \end{aligned}$$

This can be visualized as the projection, i.e. the residual $X - \mathbb{E}[X \mid Y]$ is orthogonal to the space of functions of Y, when we treat the expected value $\mathbb{E}[X \cdot Y]$ as the inner product $\langle X, Y \rangle$ between two random variables X and Y. The conditional expected value $\mathbb{E}[X \mid Y]$ is the closest point on the space of functions $h(Y)$ to X, which is proved in Problem 1.12.



Problem 1.3. Suppose that X and Θ are two random variables such that X given $\Theta = \theta$ is Poisson distributed with mean θ, i.e.,

$$f_{X|\Theta}(k|\theta) = e^{-\theta}\frac{\theta^k}{k!}, \quad k = 0, 1, 2, \dots$$

and Θ is Gamma distributed with parameters $\alpha, \beta > 0$. That is, Θ has the density function

$$f_\Theta(\theta) = \frac{\beta^\alpha \theta^{\alpha-1} e^{-\beta\theta}}{\Gamma(\alpha)}, \quad \theta > 0,$$

where Γ denotes the Gamma function,

$$\Gamma(\alpha) = \int_0^\infty \theta^{\alpha-1} e^{-\theta}\, d\theta.$$

Compute the marginal mass function of X.

Solution 1.3. The marginal mass function of X is given by

$$\begin{aligned}
\mathbb{P}(X = k) &= \int_0^\infty f_{X|\Theta}(k|\theta) f_\Theta(\theta)\, d\theta \\
&= \int_0^\infty \frac{\theta^k e^{-\theta}}{k!} \cdot \frac{\beta^\alpha \theta^{\alpha-1} e^{-\beta\theta}}{\Gamma(\alpha)}\, d\theta \\
&= \frac{\beta^\alpha}{k!\Gamma(\alpha)} \int_0^\infty \theta^{k+\alpha-1} e^{-(\beta+1)\theta}\, d\theta \\
&= \frac{\beta^\alpha}{k!\Gamma(\alpha)} \cdot \frac{1}{(\beta+1)^{k+\alpha}} \int_0^\infty x^{k+\alpha-1} e^{-x}\, dx \\
&= \frac{1}{k!\Gamma(\alpha)} \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^k \Gamma(k+\alpha) \\
&= \frac{(k+\alpha-1)(k+\alpha-2)\cdots(\alpha+1)\alpha}{k!} \left(1 - \frac{1}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^k \\
&= \binom{k+\alpha-1}{k} \left(1 - \frac{1}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^k.
\end{aligned}$$

Therefore, X follows a negative binomial distribution with parameters α and $\frac{1}{\beta+1}$.

Problem 1.4. Suppose that X given $\Theta = \theta$ is Poisson distributed with mean θ and Θ is Gamma distributed with density function

$$f_\Theta(\theta) = \frac{\beta^\alpha \theta^{\alpha-1} e^{-\beta\theta}}{\Gamma(\alpha)}, \quad \theta > 0.$$

1. Compute $\mathbb{E}[X]$.
2. Compute $\mathrm{Var}[X]$.

Recall that if $X \sim \mathrm{Poi}(\lambda)$ then $\mathbb{E}[X] = \lambda$ and $\mathrm{Var}(X) = \lambda$ and if $X \sim \Gamma(\alpha, \beta)$ then $\mathbb{E}[X] = \frac{\alpha}{\beta}$ and $\mathrm{Var}(X) = \frac{\alpha}{\beta^2}$.

Solution 1.4.

Part 1: Using the law of total expectation,

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|\Theta]] = \mathbb{E}[\Theta] = \frac{\alpha}{\beta}.$$

We used the fact that the expected value of a Poisson distributed random variable is equal to its mean parameter.

Part 2: By the law of total variance

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[\mathrm{Var}(X|\Theta)] + \mathrm{Var}(\mathbb{E}[X|\Theta]) \\ &= \mathbb{E}[\Theta] + \mathrm{Var}(\Theta) \\ &= \frac{\alpha}{\beta} + \frac{\alpha}{\beta^2} = \frac{\alpha(\beta+1)}{\beta^2}.\end{aligned}$$

We used the fact that the expected value of a Poisson distributed random variable is equal to its mean parameter and its variance is equal to its mean parameter.

Problem 1.5. Let $X \sim \mathrm{Poi}(\lambda_1)$ and $Y \sim \mathrm{Poi}(\lambda_2)$ be independent. Show that

$$X \mid X + Y = n \sim \mathrm{Bin}\left(n, \frac{\lambda_1}{\lambda_1 + \lambda_2}\right).$$

Similarly, for Y, we have

$$Y \mid X + Y = n \sim \mathrm{Bin}\left(n, \frac{\lambda_2}{\lambda_1 + \lambda_2}\right)$$

Solution 1.5. Since $X + Y \sim \mathrm{Poi}(\lambda_1 + \lambda_2)$, we have for $n \geq 0$ and $x \in 0, \dots, n$

$$\begin{aligned} p_{X \mid X+Y}(x \mid n) = \frac{p_{X,X+Y}(x,n)}{p_{X+Y}(n)} &= \frac{\mathbb{P}(X = x, X + Y = n)}{\mathbb{P}(X + Y = n)} \\ \text{independence} \quad &= \frac{\mathbb{P}(X = x)\,\mathbb{P}(Y = n - x)}{\mathbb{P}(X + Y = n)} \\ &= \frac{e^{-\lambda_1} \frac{\lambda_1^x}{x!} e^{-\lambda_2} \frac{\lambda_2^{n-x}}{(n-x)!}}{e^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1+\lambda_2)^n}{n!}} \\ &= \frac{n!}{x!(n-x)!} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^x \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{n-x} \end{aligned}$$

which we recognize as the PMF of a $\mathrm{Bin}\left(n, \frac{\lambda_1}{\lambda_1+\lambda_2}\right)$ random variable. The proof for the Y given $X + Y = n$ is identical.

Problem 1.6. Let $N \sim \mathrm{Poi}(\lambda)$ be a Poisson random variable with the mean λ. Then, consider N i.i.d. random variables, independent of N, taking values 1 or 2 with probabilities p and $q = 1 - p$ respectively. Let N_j be the number of these random variables taking value j, so that $N_1 + N_2 = N$. Show that N_1 and N_2 are independent Poisson random variables with means λp and λq respectively.

Solution 1.6. We want to compute the joint PMF of N_1 and N_2,

$$p_{N_1,N_2}(n_1, n_2) = \mathbb{P}(N_1 = n_1, N_2 = n_2) = \mathbb{P}(N_1 = n_1, N - N_1 = n_2) = \mathbb{P}(N_1 = n_1, N = n_1 + n_2).$$

By the definition of conditional probability

$$\mathbb{P}(N_1 = n_1, N = n_1 + n_2) = \mathbb{P}(N_1 = n_1 \mid N = n_1 + n_2)\,\mathbb{P}(N = n_1 + n_2).$$

By construction, when we have $N = n_1 + n_2$, then the conditional distribution of $N_1 \mid N = n_1 + n_2$ is binomial with $n_1 + n_2$ trials and probability p of success. Therefore,

$$\begin{aligned}
\mathbb{P}(N_1 = n_1 \mid N = n_1 + n_2)\,\mathbb{P}(N = n_1 + n_2) &= \binom{n_1 + n_2}{n_1} p^{n_1}(1-p)^{n_2} \cdot e^{-\lambda} \frac{\lambda^{n_1+n_2}}{(n_1+n_2)!} \\
&= \frac{(n_1+n_2)!}{n_1! n_2!} p^{n_1}(1-p)^{n_2} \cdot e^{-\lambda(p+1-p)} \frac{\lambda^{n_1+n_2}}{(n_1+n_2)!} \\
&= e^{-\lambda p} \frac{(\lambda p)^{n_1}}{n_1!} \cdot e^{-\lambda(1-p)} \frac{(\lambda(1-p))^{n_2}}{n_2!}.
\end{aligned}$$

We conclude that N_1 and N_2 are independent since it is the product of PMFs, and the marginal distribution of N_1 and N_2 are $N_1 \sim \mathrm{Poi}(\lambda p)$ and $N_2 \sim \mathrm{Poi}(\lambda(1-p))$.

Remark 4. This is called the Poisson splitting theorem and it allows you to split a Poisson process into two independent Poisson processes by randomly marking each point independently.

Problem 1.7. Suppose that

$$X = \begin{cases} \sum_{i=1}^{N} Y_i, & \text{if } N > 0, \\ 0, & \text{if } N = 0, \end{cases}$$

where N is Poisson distributed with mean λ and $Y_1, Y_2, \ldots$ is a sequence of iid random variables with mean μ and variance σ^2 that is independent of N. We say that X is a **compound Poisson random variable**.

1. Compute $\mathbb{E}[X]$.
2. Compute $\mathrm{Var}[X]$.

Recall that if $X \sim \mathrm{Poi}(\lambda)$ then $\mathbb{E}[X] = \lambda$ and $\mathrm{Var}(X) = \lambda$.

Solution 1.7.

Part 1: By the law of total expectation

$$\begin{aligned}
\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|N]] = \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{N} Y_i \mid N\right]\right] &= \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{\infty} Y_i \mathbb{1}(i \le N) \mid N\right]\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{\infty} \mathbb{1}(i \le N)\,\mathbb{E}[Y_i \mid N]\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{\infty} \mathbb{1}(i \le N)\,\mathbb{E}[Y_i]\right] \\
&= \mathbb{E}[N\mu] = \lambda\mu.
\end{aligned}$$

Part 2: By the law of total variance, and a similar computation as above

$$\begin{aligned}\mathrm{Var}\,(X) &= \mathbb{E}\left[\mathrm{Var}\,(X|N)\right] + \mathrm{Var}\,(\mathbb{E}\,[X|N]) \\ &= \mathbb{E}\left[N\sigma^2\right] + \mathrm{Var}\,(N\mu) \\ &= \sigma^2\mathbb{E}\,[N] + \mu^2\mathrm{Var}\,(N) \\ &= \lambda\left(\sigma^2+\mu^2\right).\end{aligned}$$

We used the fact that the expected value of a Poisson distributed random variable is equal to its mean parameter and its variance is equal to its mean parameter.

Remark 5. The trick with the indicators was to carefully show that we can apply linearity of expectation even when the sum is random provided that we condition on it,

$$\mathbb{E}\left[\sum_{i=1}^{N} Y_i \mid N\right] = \mathbb{E}\left[\sum_{i=1}^{\infty} Y_i \mathbb{1}(i \le N) \mid N\right] = \sum_{i=1}^{\infty} \mathbb{1}(i \le N)\,\mathbb{E}[Y_i \mid N] = \sum_{i=1}^{N} \mathbb{E}[Y_i \mid N].$$

Problem 1.8. We flip a fair coin repeatedly.

1. What is the expected number of flips until HT appears for the first time?
2. What is the expected number of flips until HH appears for the first time?

Solution 1.8.

Part 1: Let Y_{HT} denote the number of flips until the pattern HT appears for the first time. Notice that

$$Y_{HT} = Y_H + Y_T$$

where Y_H is the number of flips until the first heads and Y_T is the number of flips until the first T (after a H appeared). We can see this from the picture below

$$\underbrace{\underbrace{TTTH}_{Y_H}\underbrace{HHHHHT}_{Y_T}}_{Y_{HT}}\star\star\star$$

Since $Y_H - 1 \sim \mathrm{Geo}(0.5)$ and similarly $Y_T - 1 \sim \mathrm{Geo}(0.5)$ we have $\mathbb{E}[Y_H] = 1 + \frac{1-0.5}{0.5} = 2$ and $\mathbb{E}[Y_T] = 2$ so by the linearity of expectation

$$\mathbb{E}[Y_{HT}] = \mathbb{E}[Y_H] + \mathbb{E}[Y_T] = 2 + 2 = 4.$$

Part 2: This part is much harder to do, so we will use the law of total expectation. Let Y_{HH} denote the number of flips until we see the pattern HH for the first time and let $X_i \sim \mathrm{Ber}(0.5)$ denote the outcome of the ith flip (where 1 denotes a heads and 0 denotes a tail). We have

$$\begin{aligned}\mathbb{E}[Y_{HH}] = \mathbb{E}[\mathbb{E}[Y_{HH} \mid X_1]] &= \mathbb{E}[Y_{HH} \mid X_1 = H]\,\mathbb{P}(X_1 = H) + \mathbb{E}[Y_{HH} \mid X_1 = T]\,\mathbb{P}(X_1 = T) \\ &= \frac{1}{2}\,\mathbb{E}[Y_{HH} \mid X_1 = H] + \frac{1}{2}\,\mathbb{E}[Y_{HH} \mid X_1 = T].\end{aligned}$$

If the first coin is a tails, then we are back to square one, so we still need to find the waiting time until we get our first HH plus 1 (since we already got a T on the first flip),

$$\mathbb{E}[Y_{HH} \mid X_1 = T] = \mathbb{E}[Y_{HH}] + 1$$

If the first coin is a heads, then we are one step closer to our goal, so we can condition again on the next flip to see that

$$\begin{aligned}\mathbb{E}[Y_{HH} \mid X_1 = H] &= \mathbb{E}[\mathbb{E}[Y_{HH} \mid X_1 = H, X_2]] \\ &= \mathbb{E}[Y_{HH} \mid X_1 = H, X_2 = H]\,\mathbb{P}(X_2 = H) + \mathbb{E}[Y_{HH} \mid X_1 = H, X_2 = T]\,\mathbb{P}(X_2 = T) \\ &= \frac{1}{2}\,\mathbb{E}[Y_{HH} \mid X_1 = H, X_2 = T] + \frac{1}{2}\,\mathbb{E}[Y_{HH} \mid X_1 = H, X_2 = H].\end{aligned}$$

From here, we have that

$$\mathbb{E}[Y_{HH} \mid X_1 = H, X_2 = H] = 2$$

since we have succeeded, but

$$\mathbb{E}[Y_{HH} \mid X_1 = H, X_2 = T] = 2 + \mathbb{E}[Y_{HH}]$$

since flipping a HT means we are back to square one after 2 flips. Putting everything together implies that

$$\mathbb{E}[Y_{HH}] = \frac{1}{2}\Big(\frac{1}{2}\cdot 2 + \frac{1}{2}\Big(\mathbb{E}[Y_{HH}] + 2\Big)\Big) + \frac{1}{2}(\mathbb{E}[Y_{HH}] + 1) \implies \mathbb{E}[Y_{HH}] = \frac{3}{4}\,\mathbb{E}[Y_{HH}] + \frac{3}{2}$$

which after some algebra gives $\mathbb{E}[Y_{HH}] = 6$.

Remark 6. Somewhat surprisingly the expected values are different. The main difference is that the time to get the first HT is that after we flipped our first heads, we have already made progress towards our goal. However, to get HH if we flip the first heads, then flip a tails, we are back to square one and lost the partial progress. This is obvious if we solve Part 1 using the logic of Part 2,

$$\mathbb{E}[Y_{HT}] = \mathbb{E}[\mathbb{E}[Y_{HT} \mid X_1]] = \frac{1}{2}\,\mathbb{E}[Y_{HT} \mid X_1 = H] + \frac{1}{2}\,\mathbb{E}[Y_{HT} \mid X_1 = T]$$

We have that $\mathbb{E}[Y_{HT} \mid X_1 = H] = 1 + \mathbb{E}[Y_T] = 3$ and $\mathbb{E}[Y_{HT} \mid X_1 = T] = \mathbb{E}[Y_{HT}] + 1$ so

$$\mathbb{E}[Y_{HT}] = \frac{1}{2}\cdot 3 + \frac{1}{2}(\mathbb{E}[Y_{HT}] + 1) \implies \mathbb{E}[Y_{HT}] = 4.$$

Problem 1.9.

1. Let $X \sim \mathrm{Unif}[-1,1]$ and $Y = |X|$. Compute
$$\mathbb{E}[X \mid Y]$$

2. Let $X \sim \mathrm{Unif}[-1,2]$ and $Y = |X|$. Compute
$$\mathbb{E}[X \mid Y]$$

3. Let $X \sim \mathrm{Unif}[-1,1]$ and $Y = |X|$. Compute
$$\mathbb{E}[X^2 \mid Y]$$

Solution 1.9. Even though X and Y are continuous random variables, we cannot define the joint density because (X,Y) is supported on the line $y = |x|$, which is a 1 dimensional curve and therefore has no area.

Intuitively, if we know $Y = |X| = c$, then we would expect that either $X = c$ or $X = -c$ with the same probability, since X is symmetric and can only take two values if we know its absolute value. We will have to use the abstract definition of the conditional expectation to make this intuition precise.

Since we will reuse this in Part 2, for now let's suppose that $f_X(x)$ is arbitrary. We let $A \subset [0,\infty)$ be a Borel set (since it suffices to take a Borel set contained in the range of Y)

$$\begin{aligned}
\mathbb{E}[X\mathbb{1}(|X|\in A)] &= \int_{-\infty}^{\infty} x\mathbb{1}(|x|\in A) f_X(x)\,dx \\
&= \int_0^{\infty} x\mathbb{1}(|x|\in A) f_X(x)\,dx + \int_{-\infty}^{0} x\mathbb{1}(|x|\in A) f_X(x)\,dx \\
&\overset{y=-x}{=} \int_0^{\infty} x\mathbb{1}(x\in A) f_X(x)\,dx + \int_0^{\infty} (-y)\mathbb{1}(y\in A) f_X(-y)\,dy \\
&= \int_A x(f_X(x) - f_X(-x))\,dx.
\end{aligned}$$

Our goal is to find a function $g(Y) = \mathbb{E}[X \mid Y]$ so that

$$\mathbb{E}[g(Y)\mathbb{1}(Y\in A)] = \int_A g(y) f_Y(y)\,dy = \int_A x(f_X(x) - f_X(-x))\,dx.$$

Since $F_Y(y) = \mathbb{P}(|X| \le y) = \mathbb{P}(-y \le X \le y) = F_X(y) - F_X(-y)$ we have $f_Y(y) = f_X(y) + f_X(-y)$. Therefore, we must take

$$g(y) = \begin{cases} y\frac{f_X(y)-f_X(-y)}{f_X(y)+f_X(-y)} & y \in \mathcal{Y} \\ 0 & \text{otherwise} \end{cases}.$$

Part 1: If $X \sim \mathrm{Unif}[-1,1]$, then we must have for $y \in \mathcal{Y}$

$$g(y) = y\frac{f_X(y)-f_X(-y)}{f_X(y)+f_X(-y)} = y\frac{\frac{1}{2}-\frac{1}{2}}{\frac{1}{2}+\frac{1}{2}} = 0$$

so

$$\mathbb{E}[X \mid Y] = g(Y) = 0.$$

This is intuitive because the best guess of X given $|X|$ is 0 by symmetry since you won't prefer $+|X|$ or $-|X|$.

Part 2: If $X \sim \mathrm{Unif}[-1,2]$, then we have different cases. For $y \in [1,2]$, we have

$$g(y) = y\frac{f_X(y)-f_X(-y)}{f_X(y)+f_X(-y)} = y\frac{\frac{1}{3}-0}{\frac{1}{3}+0} = y$$

and for $y \in [0,1]$

$$g(y) = y\frac{f_X(y)-f_X(-y)}{f_X(y)+f_X(-y)} = y\frac{\frac{1}{3}-\frac{1}{3}}{\frac{1}{3}+\frac{1}{3}} = 0$$

so

$$\mathbb{E}[X \mid Y] = Y\mathbb{1}(Y \in [1,2]).$$

This is intuitive because the best guess of X given $|X| \in [0,1]$ is 0 by symmetry since you won't prefer $+|X|$ or $-|X|$. However, if given that $|X| = y$ for $y > 1$, then our best guess for X is y since we know that $-y$ is not possible.

Part 3: A simple modification of the above argument with X^2 instead of X implies that

$$\begin{aligned}\mathbb{E}[X^2\mathbb{1}(|X|\in A)] &= \int_{-\infty}^{\infty} x^2\mathbb{1}(|x|\in A)f_X(x)\,dx \\ &= \int_0^{\infty} x^2\mathbb{1}(|x|\in A)f_X(x)\,dx + \int_{-\infty}^{0} x^2\mathbb{1}(|x|\in A)f_X(x)\,dx \\ &\overset{y=-x}{=} \int_0^{\infty} x^2\mathbb{1}(x\in A)f_X(x)\,dx + \int_0^{\infty} (-y)^2\mathbb{1}(y\in A)f_X(-y)\,dy \\ &= \int_A x^2(f_X(x)+f_X(-x))\,dx.\end{aligned}$$

Our goal is to find a function $g(Y)=\mathbb{E}[X^2 \mid Y]$ so that

$$\mathbb{E}[g(Y)\mathbb{1}(Y\in A)] = \int_A g(y)f_Y(y)\,dy = \int_A x^2(f_X(x)+f_X(-x))\,dx.$$

Since $f_Y(y)=f_X(y)+f_X(-y)$ we must have

$$g(y) = \begin{cases} y^2\frac{f_X(y)+f_X(-y)}{f_X(y)+f_X(-y)} & y\in\mathcal{Y} \\ 0 & \text{otherwise}\end{cases}.$$

If $X\sim \mathrm{Unif}[-1,1]$, then we must have for $y\in\mathcal{Y}$

$$g(y) = y^2\frac{f_X(y)+f_X(-y)}{f_X(y)+f_X(-y)} = y^2\frac{\frac{1}{2}+\frac{1}{2}}{\frac{1}{2}+\frac{1}{2}} = y^2$$

so

$$\mathbb{E}[X^2 \mid Y] = g(Y) = Y^2.$$

This is intuitive because the best guess of X^2 given $|X|$ is simply $|X|^2$ since we know $|X|$ completely determines $|X|^2$.

1.5 Proofs of Key Results

Problem 1.10. Prove the properties of conditional expectations in Proposition 1.

Solution 1.10. The properties follow directly from the definition. We state it for continuous conditional distributions, but the same proofs hold for discrete ones.

Independence: If X and Y are independent, then $f_{X\mid Y}(x\mid y)=f_X(x)$, so by the definition,

$$\mathbb{E}[X\mid Y] = \int_{\mathbb{R}} x f_{X|Y}(x|Y)\,dx = \int_{\mathbb{R}} x f_X(x)\,dx = \mathbb{E}[X].$$

Linearity: Linearity follows from the fact that integrals are linear

$$\begin{aligned}\mathbb{E}[X_1+X_2\mid Y=y] &= \iint (x_1+x_2)f_{X_1,X_2\mid Y}(x_1,x_2\mid Y=y)\,dx_1dx_2 \\ &= \iint x_1 f_{X_1,X_2\mid Y}(x_1,x_2\mid Y=y)\,dx_1dx_2 + \iint x_2 f_{X_1,X_2\mid Y}(x_1,x_2\mid Y=y)\,dx_1dx_2 \\ &= \iint x_1\frac{f_{X_1,X_2,Y}(x_1,x_2,y)}{f_Y(y)}\,dx_2dx_1 + \iint x_2\frac{f_{X_1,X_2,Y}(x_1,x_2,y)}{f_Y(y)}\,dx_1dx_2 \\ &= \int x_1\frac{f_{X_1,Y}(x_1,y)}{f_Y(y)}\,dx_1 + \int x_2\frac{f_{X_2,Y}(x_2,y)}{f_Y(y)}\,dx_2 \\ &= \mathbb{E}[X_1\mid Y=y]+\mathbb{E}[X_2\mid Y=y].\end{aligned}$$

Pulling out Known Factors: For any y in the support of Y,

$$g(y) = \mathbb{E}\left[h(Y)X|Y=y\right] = \int_{\mathbb{R}} h(y)xf_{X|Y}(x|y)\ dx = h(y)\int_{\mathbb{R}} xf_{X|Y}(x|y)\ dx = h(y)\mathbb{E}\left[X|Y=y\right].$$

Therefore,

$$\mathbb{E}\left[h(Y)X|Y\right] = g(Y) = h(Y)\mathbb{E}\left[X|Y\right].$$

Law of Total Expectation: We define $g(y) = \mathbb{E}\left[X|Y=y\right] = \int_{\mathbb{R}} xf_{X|Y}(x|y)\,dx$. By the definition of the expected value,

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[X|Y\right]\right] = \mathbb{E}\left[g(Y)\right] &= \int_{\mathbb{R}} g(y)f_Y(y)\,dy = \int_{\mathbb{R}}\left(\int_{\mathbb{R}} xf_{X|Y}(x|y)\,dx\right)f_Y(y)\,dy\\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} xf_{X|Y}(x|y)\,f_Y(y)\,dx\,dy\\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} xf_{X,Y}(x,y)\ dx\,dy\\
&= \int_{\mathbb{R}} x\left(\int_{\mathbb{R}} f_{X,Y}(x,y)\ dy\right)\ dx\\
&= \int_{\mathbb{R}} xf_X(x)\,dx = \mathbb{E}\left[X\right].
\end{aligned}$$

Problem 1.11. Suppose that $\mathbb{E}[X^2] < \infty$. For any constant c, show that

$$\mathbb{E}[(X-c)^2] \geq \mathbb{E}[(X-\mathbb{E}[X])^2].$$

In particular, the expected value is the constant that minimizes the mean squared error.

Solution 1.11. This proof follows directly from the properties of the expected value. By adding and subtracting $\mathbb{E}[X]$, we see that

$$\begin{aligned}
\mathbb{E}[(X-c)^2] &= \mathbb{E}[(X-\mathbb{E}[X]+\mathbb{E}[X]-c)^2]\\
&= \mathbb{E}[(X-\mathbb{E}[X])^2] + \mathbb{E}[(\mathbb{E}[X]-c)^2] + 2\,\mathbb{E}[(X-\mathbb{E}[X])(\mathbb{E}[X]-c)]
\end{aligned}$$

Since $\mathbb{E}[X]-c$ is not random, we see that the cross terms vanish

$$\mathbb{E}[(X-\mathbb{E}[X])(\mathbb{E}[X]-c)] = (\mathbb{E}[X]-c)\,\mathbb{E}[(X-\mathbb{E}[X])] = (\mathbb{E}[X]-c)(\mathbb{E}[X]-\mathbb{E}[X]) = 0.$$

Since $\mathbb{E}[(\mathbb{E}[X]-c)^2] \geq 0$, we conclude that

$$\mathbb{E}[(X-c)^2] = \mathbb{E}[(X-\mathbb{E}[X])^2] + \mathbb{E}[(\mathbb{E}[X]-c)^2] \geq \mathbb{E}[(X-\mathbb{E}[X])^2]$$

as required.

Problem 1.12. Suppose that $\mathbb{E}[X^2] < \infty$. For any measurable function f, show that

$$\mathbb{E}[(X-f(Y))^2] \geq \mathbb{E}[(X-\mathbb{E}[X \mid Y])^2].$$

In particular, the conditional expectation minimizes the mean squared error.

Solution 1.12. This proof follows directly from the properties of the conditional expected value. By adding and subtracting $\mathbb{E}[X \mid Y]$, we see that

$$\begin{aligned}\mathbb{E}[(X - f(Y))^2] &= \mathbb{E}[(X - \mathbb{E}[X \mid Y] + \mathbb{E}[X \mid Y] - f(Y))^2] \\ &= \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2] + \mathbb{E}[(\mathbb{E}[X \mid Y] - f(Y))^2] + 2\,\mathbb{E}[(X - \mathbb{E}[X \mid Y])(\mathbb{E}[X \mid Y] - f(Y))]\end{aligned}$$

Applying the law of total expectation and using the fact that $\mathbb{E}[X \mid Y]$ and $f(Y)$ are measurable functions of Y, we see that the cross terms vanish

$$\begin{aligned}\mathbb{E}[(X - \mathbb{E}[X \mid Y])(\mathbb{E}[X \mid Y] - f(Y))] &= \mathbb{E}\left[\mathbb{E}[(X - \mathbb{E}[X \mid Y])(\mathbb{E}[X \mid Y] - f(Y)) \mid Y]\right] \\ &= \mathbb{E}\left[(\mathbb{E}[X \mid Y] - f(Y))\,\mathbb{E}[(X - \mathbb{E}[X \mid Y]) \mid Y]\right] \\ &= \mathbb{E}[(\mathbb{E}[X \mid Y] - f(Y))(\mathbb{E}[X \mid Y] - \mathbb{E}[X \mid Y])] \\ &= 0.\end{aligned}$$

Since $\mathbb{E}[(\mathbb{E}[X \mid Y] - f(Y))^2] \geq 0$, we conclude that

$$\mathbb{E}[(X - f(Y))^2] = \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2] + \mathbb{E}[(\mathbb{E}[X \mid Y] - f(Y))^2] \geq \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2]$$

as required.

Remark 7. Notice that this proof only uses the law of total expectation and the trick that allows us to pull known factors.

Remark 8. Problem 1.11 is a special case when we look at estimators that do not depend on Y.

Remark 9. If X was a function of Y, say $h(Y)$, then the result says that the "best" estimate (in terms of squared error) for X given information Y is $h(Y)$,

$$\mathbb{E}[X \mid Y] = \mathbb{E}[h(Y) \mid Y] = h(Y)\,\mathbb{E}[1 \mid Y] = h(Y).$$

We have that $\mathbb{E}[X \mid Y]$ is still the "best" estimate even when X can be some complicated possibly random function of Y.

Problem 1.13. Prove the properties of conditional variance in Proposition 2.

Solution 1.13.

Part 1: With $g(Y) = \mathbb{E}[X|Y]$ we have from Proposition 1 part 3 that

$$\begin{aligned}\mathrm{Var}\,(X|Y) &= \mathbb{E}\left[X^2 - 2X\mathbb{E}[X|Y] + (\mathbb{E}[X|Y])^2 \,\middle|\, Y\right] \\ &= \mathbb{E}\left[X^2 \mid Y\right] - 2\mathbb{E}\left[X\mathbb{E}[X|Y] \mid Y\right] + \mathbb{E}\left[(\mathbb{E}[X|Y])^2 \middle| Y\right] \\ &= \mathbb{E}\left[X^2 \mid Y\right] - 2\mathbb{E}\left[Xg(Y) \mid Y\right] + \mathbb{E}\left[(g(Y))^2 \middle| Y\right] \\ &= \mathbb{E}\left[X^2 \mid Y\right] - 2g(Y) \cdot \mathbb{E}\left[X \mid Y\right] + (g(Y))^2\mathbb{E}[1|Y] \qquad \text{(by Proposition 1 part 3)} \\ &= \mathbb{E}\left[X^2 \mid Y\right] - 2\mathbb{E}\left[X \mid Y\right] \cdot \mathbb{E}\left[X \mid Y\right] + (\mathbb{E}[X|Y])^2 \\ &= \mathbb{E}\left[X^2 \mid Y\right] - (\mathbb{E}\left[X \mid Y\right])^2\end{aligned}$$

Part 2: It follows from Part 1 and Proposition 1 Part 4 that

$$\begin{aligned}\mathbb{E}\left[\mathrm{Var}\,(X|Y)\right] &= \mathbb{E}\left[\mathbb{E}\left[X^2|Y\right]\right] - \mathbb{E}\left[(\mathbb{E}\left[X|Y\right])^2\right] \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}\left[(\mathbb{E}\left[X|Y\right])^2\right].\end{aligned}$$

On the other hand,

$$\begin{aligned}\operatorname{Var}\left(\mathbb{E}\left[X|Y\right]\right) &= \mathbb{E}\left[\left(\mathbb{E}\left[X|Y\right]\right)^2\right] - \left(\mathbb{E}\left[\mathbb{E}\left[X|Y\right]\right]\right)^2 \\ &= \mathbb{E}\left[\left(\mathbb{E}\left[X|Y\right]\right)^2\right] - \left(\mathbb{E}\left[X\right]\right)^2.\end{aligned}$$

Combining the preceding two relations implies

$$\mathbb{E}\left[\operatorname{Var}\left(X|Y\right)\right] + \operatorname{Var}\left(\mathbb{E}\left[X|Y\right]\right) = \mathbb{E}\left[X^2\right] - \left(\mathbb{E}\left[X\right]\right)^2 = \operatorname{Var}\left(X\right).$$

Problem 1.14. Show that the conditional expectation is unique, i.e. there is only one function $g(Y) = \mathbb{E}[X|Y]$ that satisfies

$$\mathbb{E}\big[g(Y)\mathbb{1}(Y \in A)\big] = \mathbb{E}[X\mathbb{1}(Y \in A)]$$

for all Borel sets $A \in \mathbb{R}$.

Solution 1.14. Suppose there exists functions $g(Y) = \mathbb{E}[X \mid Y]$ and $h(Y) = \mathbb{E}[X \mid Y]$ that are distinct in the sense that $\mathbb{P}(g(Y) \neq h(Y)) > 0$. We must have that either $\mathbb{P}(g(Y) > h(Y)) > 0$ or $\mathbb{P}(g(Y) < h(Y)) > 0$. Suppose without loss of generality that the first holds. Since both $g(Y)$ and $h(Y)$ are measurable, the event $A = \{g(y) > h(y)\}$ is a Borel set and $\mathbb{P}(g(Y) - h(Y) > 0) = \mathbb{P}(g(Y) > h(Y)) > 0$, so

$$\mathbb{E}[(g(Y) - h(Y))\mathbb{1}(Y \in A)] > 0$$

since $g(Y) - h(Y) > 0$ on A and $\mathbb{P}(Y \in A) > 0$. However, by the definition of conditional expectations $\mathbb{E}[g(Y)\mathbb{1}(Y \in A)] = \mathbb{E}[h(Y)\mathbb{1}(Y \in A)] = \mathbb{E}[X\mathbb{1}(Y \in A)]$, so

$$\mathbb{E}[(g(Y) - h(Y))\mathbb{1}(Y \in A)] = \mathbb{E}[g(Y)\mathbb{1}(Y \in A)] - \mathbb{E}[h(Y)\mathbb{1}(Y \in A)] = \mathbb{E}[X\mathbb{1}(Y \in A)] - \mathbb{E}[X\mathbb{1}(Y \in A)] = 0$$

which is a contradiction.

Problem 1.15. Prove all the statements in Proposition 1 using the abstract definition of conditional expectations.

Solution 1.15. These properties will follow from manipulations of the definition of conditional expectation.

1. **Independence:** $\mathbb{E}[X]$ is trivially a function of Y because it is a constant function. We show that $\mathbb{E}[X]$ satisfies the definition of conditional expectation of $\mathbb{E}[X \mid Y]$, i.e. for any Borel set $A \in \mathbb{R}$ we have

$$\mathbb{E}[\mathbb{E}[X]\mathbb{1}(Y \in A)] = \mathbb{E}[X]\,\mathbb{E}[\mathbb{1}(Y \in A)] \underbrace{=}_{\star} \mathbb{E}[X\mathbb{1}(Y \in A)] = \mathbb{E}[\mathbb{E}[X \mid Y]\mathbb{1}(Y \in A)]$$

where in $\star$ we used the fact that if X and Y are independent then $\mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}[XY]$. We conclude that $\mathbb{E}[X] = \mathbb{E}[X \mid Y]$ by uniqueness.

2. **Linearity:** This implies that the conditional expectation behaves like the usual expectation. We check that the right hand side satisfies the definition of conditional expectation for the random variable $\mathbb{E}[aX_1 + bX_2|Y]$, i.e. for any Borel set $A \subset \mathbb{R}$

$$\begin{aligned}\mathbb{E}[(a\,\mathbb{E}[X_1 \mid Y] + b\,\mathbb{E}[X_2 \mid Y])\mathbb{1}(Y \in A)] &= a\,\mathbb{E}[\mathbb{E}[X_1 \mid Y]\mathbb{1}(Y \in A)] + b\,\mathbb{E}[\mathbb{E}[X_2 \mid Y]\mathbb{1}(Y \in A)] \\ &= a\,\mathbb{E}[X_1\mathbb{1}(Y \in A)] + b\,\mathbb{E}[X_2\mathbb{1}(Y \in A)] \\ &= \mathbb{E}[(aX_1 + bX_2)\mathbb{1}(Y \in A)] \\ &= \mathbb{E}[\mathbb{E}[(aX_1 + bX_2) \mid Y]\mathbb{1}(Y \in A)].\end{aligned}$$

Therefore, by uniqueness $\mathbb{E}[aX_1 + bX_2 \mid Y] = a\,\mathbb{E}[X_1 \mid Y] + b\,\mathbb{E}[X_2 \mid Y]$.

3. **Pulling out known factors:** The statement itself is very natural, since $h(Y)$ is in some sense not random once given information Y, so we can factor out.

   The proof is a bit technical for this course, but we will include it here for completeness. It suffices to show this result for simple functions $h(Y) = \mathbb{1}(Y \in B)$, because more general functions can be approximated by linear combinations of such random variables, so the result will follow from linearity. Therefore, it remains to show that

$$\mathbb{E}[\mathbb{1}(Y \in B)X \mid Y] = \mathbb{1}(Y \in B)\,\mathbb{E}[X \mid Y].$$

   Notice that for any Borel set $A \subset \mathbb{R}$, $A \cap B \subset \mathbb{R}$ is a Borel set by the definition of a σ-algebra. We check that the random variable $\mathbb{1}(Y \in B)\,\mathbb{E}[X \mid Y]$ (which is clearly a function of Y) satisfies the definition of the conditional expectation $\mathbb{E}[\mathbb{1}(Y \in B)X \mid Y]$, i.e. for any Borel set A

$$\begin{aligned}\mathbb{E}[(\mathbb{1}(Y \in B)\,\mathbb{E}[X \mid Y])\mathbb{1}(Y \in A)] &= \mathbb{E}[\mathbb{E}[X \mid Y]\mathbb{1}(Y \in A \cap B)] \\ &= \mathbb{E}[X\mathbb{1}(Y \in A \cap B)] \\ &= \mathbb{E}[(X\mathbb{1}(Y \in B))\mathbb{1}(Y \in A)] = \mathbb{E}[\mathbb{E}[\mathbb{1}(Y \in B)X \mid Y]\mathbb{1}(Y \in A)].\end{aligned}$$

   By uniqueness, $\mathbb{E}[\mathbb{1}(Y \in B)X \mid Y] = \mathbb{1}(Y \in B)\,\mathbb{E}[X \mid Y]$.

4. **Law of Total Expectation:** Taking Borel set $A = \mathbb{R}$, we have that

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[\mathbb{E}[X \mid Y]\mathbb{1}(Y \in \mathbb{R})] = \mathbb{E}[X\mathbb{1}(Y \in \mathbb{R})] = \mathbb{E}[X].$$