1 Conditional Expectation

We saw that conditional probabilities represent how our beliefs for probabilities of events 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 Review

Recall the following concepts we defined in Week 8,

- the **marginal mass or density function of** X

$$p_X(x) = \sum_y p_{X,Y}(x,y) \qquad \text{or} \qquad f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)\, dy.$$

- the **marginal mass or density function of** Y

$$p_Y(y) = \sum_x p_{X,Y}(x,y) \qquad \text{or} \qquad f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x,y)\, dx.$$

- the **conditional mass or density function of** X **given** $Y = y$

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \quad \text{or} \quad f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}. \tag{1}$$

  provided that we are not dividing by zero.

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 1 (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 1.9). 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 second 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 2 (Conditional Variance)**.** The *conditional variance* of X given Y is defined as
$$\mathrm{Var}\left(X|Y\right) = \mathbb{E}\left[(X - \mathbb{E}[X|Y])^2|Y\right]$$

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

1.3 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 $Y = X^2$. Writing down the conditional PDF is possible, but the intuition is very clear, and we don't have to compute it. 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, it will use techniques we have not learned in this course since the joint distribution is neither discrete nor continuous.

Problem 1.2. For any function h, show that

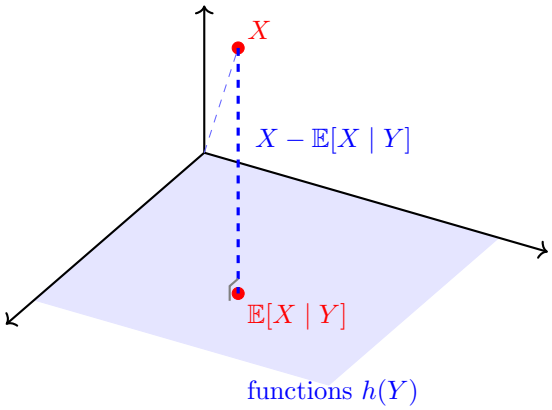
$$\mathbb{E}[(X - \mathbb{E}[X \mid Y])h(Y)] = 0.$$

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 line residual $Y - \mathbb{E}[Y \mid X]$ is orthogonal to the space of functions of X, 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 $\mathbb{E}[Y \mid X]$ is the closest point on the space of functions $h(Y)$ to X, which is proved in Problem 1.9.



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.,

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

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. 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 $\text{Var}[X]$.

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

Solution 1.5.

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

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

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 4. 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 1.6. 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 HT appears for the first time?

Solution 1.6.

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

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

Since $Y_H - 1 \sim \text{Geo}(0.5)$ and $Y_T - 1 \sim \text{Geo}(0.5)$ we have $\mathbb{E}[Y_H] = \frac{1}{0.5} = 2$ and $\mathbb{E}[Y_T] = \frac{1}{0.5} = 2$ 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 $X_i \sim \text{Ber}(0.5)$ denote the outcome of the ith flip (where 1 denotes a heads and 0 denotes a tail). We have

$$\begin{aligned}\mathbb{E}[Y_{HH}] = \mathbb{E}[\mathbb{E}[Y_{HH} \mid X_1]] &= \mathbb{E}[Y_{HH} \mid X_1 = 1]\,\mathbb{P}(X_1 = 1) + \mathbb{E}[Y_{HH} \mid X_1 = 0]\,\mathbb{P}(X_1 = 0) \\ &= \frac{1}{2}\,\mathbb{E}[Y_{HH} \mid X_1 = 1] + \frac{1}{2}\,\mathbb{E}[Y_{HH} \mid X_1 = 0].\end{aligned}$$

If the first coin is a tails, then we are back to square one, so the 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),

$$\mathbb{E}[Y_{HH} \mid X_1 = 0] = \mathbb{E}[Y_{HH}] + 1$$

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

$$\begin{aligned}\mathbb{E}[Y_{HH} \mid X_1 = 0] &= \mathbb{E}[\mathbb{E}[Y_{HH} \mid X_1 = 1, X_2]] \\ &= \mathbb{E}[Y_{HH} \mid X_1 = 1, X_2 = 1]\,\mathbb{P}(X_2 = 0) + \mathbb{E}[Y_{HH} \mid X_1 = 1, X_2 = 0]\,\mathbb{P}(X_2 = 1) \\ &= \frac{1}{2}\,\mathbb{E}[Y_{HH} \mid X_1 = 0, X_2 = 1] + \frac{1}{2}\,\mathbb{E}[Y_{HH} \mid X_1 = 0, X_2 = 0].\end{aligned}$$

From here, we have that

$$\mathbb{E}[Y_{HH} \mid X_1 = 1, X_2 = 1] = 2$$

since we have succeeded, but

$$\mathbb{E}[Y_{HH} \mid X_1 = 0, X_2 = 0] = 2 + \mathbb{E}[Y_{HH}]$$

since flipping a HT means we hare 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 5. 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,

$$\mathbb{E}[Y_{HT}] = \mathbb{E}[\mathbb{E}[Y_{HT} \mid X_1]] = \frac{1}{2}\,\mathbb{E}[Y_{HT} \mid X_1 = 1] + \frac{1}{2}\,\mathbb{E}[Y_{HT} \mid X_1 = 0]$$

We have that $\mathbb{E}[Y_{HT} \mid X_1 = 1] = \mathbb{E}[Y_T] = 2$ and $\mathbb{E}[Y_{HT} \mid X_1 = 0] = \mathbb{E}[Y_{HT}] + 1$ so

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

1.4 Proofs of Key Results

Problem 1.7. Prove the properties of conditional expectations in Proposition 1.

Solution 1.7. The properties follow directly from the definition.

Part 1: If X and Y are independent, then $f_{X \mid Y}(x \mid y) = f_X(x)$, so by the definition,

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

Part 2: Linearity follows from the fact that integrals are linear and $f_{X_1+X_2 \mid Y}(x_1 + x_2 \mid y)$ is a PDF for each fixed y.

Part 3: For any y in the support of Y,

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

Therefore,

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

Part 4: We define $g(y) = \mathbb{E}\left[X|Y = y\right] = \int_{\mathbb{R}} x f_{X|Y}\,(x|y)\,dx$. By the definition of the expected value,

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

Problem 1.8. 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 1.8. 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 1.9. 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 1.9. 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 6. Notice that this proof only uses the law of total expectation and the trick that allows us to pull known factors.

Remark 7. Problem 1.9 is a special case when we look at estimators that do not depend on Y.

Remark 8. 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 1.10. Prove the properties of conditional variance in Proposition 2.

Solution 1.10.

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

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

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

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

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

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