Suppose that a frog is standing on a lily pad, which is in the middle of a long line of lily pads. a lily pads to her right is her house, which she is trying to get to; however, it's dark and she's lost, so she decides to try to get home by randomly jumping left or right at each step, where each direction of jump is equally likely. However, b steps to the left of her initial position is a missing lily pad, so if she tries to jump to that space, she will fall into the water. A natural question to ask is what is her probability of getting home safely.

In order to approach this problem, let's reframe it completely. In this scenario, there is a casino where the only game is the coin-flipping game: every turn you choose to play, a fair coin is tossed, and if it comes up heads, you gain \$1, whereas if it comes up tails, you lose \$1. Since this is a fair game, our intuition tells us that no strategy you pick can make it so you win, on average, a positive amount of money; if such a strategy existed, real-life casinos couldn't operate, since everyone would come in every day and make a positive amount of money, on average.

That said, let's consider some strategies you could invoke. For instance, one strategy might be to play for exactly one round, and for this strategy we can see directly that your average payoff is \$0. A more complex strategy is to keep playing until you've either gained \$$a$ or lost \$$b$. Let p_W be the probability that, when you use this strategy, you end up winning \$$a$, and $p_L = 1 - p_W$ be the probability of losing \$$b$. Again, our intuition tells us that your average winnings should be \$0, which is the same as saying that

$$0 = p_W \cdot a + p_L \cdot (-b).$$

Plugging in $p_L = 1 - p_W$ gives

$$0 = p_W \cdot a + (1 - p_W)(-b) = p_W(a + b) - b.$$

and rearranging gives us

$$p_W = \frac{b}{a+b}, \qquad p_L = \frac{a}{a+b}.$$

Now, how is this related to our poor lost frog? Well, we can think of each random choice she makes as a fair coin toss, where she advances one lily pad to the right if it comes up heads, and one to the left if not. In this case, her total winnings are exactly the same as the amount she has progressed towards home, so in other words these two problems are completely equivalent, and we find that her probability of getting home is exactly $p_W = b/(a+b)$.

However, there's something problematic about this argument, which is that we haven't proved our intuitive idea that no strategy can win a fair game. And this is a big problem: the blurb for this class describes a different strategy for essentially the same fair game, where the average amount won is positive! So in order to make the above computation rigorous (along with many other related computations, for instance the average time it'll take the frog to either get home or fall in the water), we will develop the theory of martingales and understand the conditions that make fair games actually unbeatable.

1 Random Variables

Before we can define martingales, we first need to get comfortable with random variables and their properties.

Definition 1.1. A random variable X is a variable which can take on one of several values, each with some probability. We will sometimes denote the set of values that X can take by $\mathcal{X}$, which is just some non-empty subset of $\mathbb{R}$. $\mathcal{X}$ is often called the *state space* of X.

As always, when dealing with random events, the probability that something happens is a real number between 0 and 1, and if we add up the probabilities over all possible outcomes, we need to get 1.

Example 1.2.

1. Suppose we roll a fair die, and let X be the outcome. Then $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$, and
$$\Pr(X = 1) = \Pr(X = 2) = \Pr(X = 3) = \Pr(X = 4) = \Pr(X = 5) = \Pr(X = 6) = \frac{1}{6}.$$

2. Suppose we flip three fair coins, and let H be the number of heads that come up. Then $\mathcal{H} = \{0, 1, 2, 3\}$, and we can calculate the probability that H takes each of these values. First, $\Pr(H = 0)$ is the probability that none of the coins come up heads, namely that they all come up tails. Since there is a $1/2$ chance of this happening for each coin, we see that
$$\Pr(H = 0) = \left(\frac{1}{2}\right)^3 = \frac{1}{8}.$$
Similarly, $\Pr(H = 3) = 1/8$, since this is the probability that all three coins come up heads. For $\Pr(H = 1)$, observe that we need one coin to come up heads and two to come up tails; there are three choices for the heads coin, and once we've made that choice, the probability of the configuration is again $1/8$. In other words,
$$\Pr(H = 1) = \frac{3}{8}.$$
We can use a similar argument to determine that $\Pr(H = 2) = 3/8$ as well, though we can also find this out by recalling that the sum of the probabilities must be 1, so
$$\Pr(H = 2) = 1 - \Pr(H = 0) - \Pr(H = 1) - \Pr(H = 3) = 1 - \frac{1}{8} - \frac{3}{8} - \frac{1}{8} = \frac{3}{8}.$$

3. Let X again be the outcome of a roll of a fair die. We define two new random variables:
$$Y = \begin{cases} 1 & \text{if } X \text{ is prime,} \\ 0 & \text{otherwise.} \end{cases} \qquad Z = \begin{cases} 1 & \text{if } X \text{ is even,} \\ 0 & \text{otherwise.} \end{cases}$$

Such random variables are often called *indicator variables*, since their value indicates whether an event has taken place (their value is 1 if it has, 0 if it has not). Then we have that $\mathcal{Y} = \mathcal{Z} = \{0, 1\}$, and we can calculate that

$$\Pr(Y = 1) = \Pr(X \text{ is prime}) = \Pr(X = 2) + \Pr(X = 3) + \Pr(X = 5) = \frac{1}{2},$$
$$\Pr(Z = 1) = \Pr(X \text{ is even}) = \Pr(X = 2) + \Pr(X = 4) + \Pr(X = 6) = \frac{1}{2}.$$

This shows that both Y and Z behave like the number of heads when we toss a fair coin, namely they both take the values 0 and 1, each with probability 1/2. In the previous example, we implicitly used the idea that when we toss two fair coins, the probability that they both come up heads is $(1/2)^2 = 1/4$; however, this is no longer the case for Y and Z:

$$\Pr(Y = 1 \text{ and } Z = 1) = \Pr(X \text{ is prime and even}) = \Pr(X = 2) = \frac{1}{6} \neq \frac{1}{4}.$$

1.1 Dependence, independence, and conditional probability

The phenomenon exhibited in the last example is *dependence*, which basically is just the fact that different random variables can interact with one another. It is not sufficient to just know the distribution of each random variable (i.e. which values it takes and with what probabilities); we need to also know how these distributions can affect one another.

Definition 1.3. Two random variables Y and Z are called *independent* if, for all $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$, we have

$$\Pr(Y = y \text{ and } Z = z) = \Pr(Y = y) \Pr(Z = z).$$

They are called *dependent* if they are not independent.

The way we understand dependence is with *conditional probability*, which measures the probability that Y takes the value y, given that we know that some other variable Z takes on the value z. To get an intuition for what this quantity should be, suppose we write the following table: the rows are labeled by elements of $\mathcal{Y}$, the columns are labeled by elements of $\mathcal{Z}$, and in the entry corresponding to $y_i \in \mathcal{Y}, z_j \in \mathcal{Z}$, we put the probability $\Pr(Y = y_i \text{ and } Z = z_j)$:

$\mathcal{Y}$ \ $\mathcal{Z}$	z_1	z_2	$\cdots$	z_m
y_1	$\Pr(Y = y_1, Z = z_1)$	$\Pr(Y = y_1, Z = z_2)$	$\cdots$	$\Pr(Y = y_1, Z = z_m)$
y_2	$\Pr(Y = y_2, Z = z_1)$	$\Pr(Y = y_2, Z = z_2)$	$\cdots$	$\Pr(Y = y_2, Z = z_m)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
y_n	$\Pr(Y = y_n, Z = z_1)$	$\Pr(Y = y_n, Z = z_2)$	$\cdots$	$\Pr(Y = y_n, Z = z_m)$

This table has some nice properties; for instance, the sum of the entries in the ith row is exactly $\Pr(Y = y_i)$, and the sum of the entries in the jth column is $\Pr(Z = z_j)$. Indeed, by summing across the row or column, we are taking into account all possible values of the other variable, leaving us only the probability of one of the two variables.

Now, suppose we are told that $Z = z_j$ for some $z_j \in \mathcal{Z}$. This means that to calculate the conditional probabilities of Y, we should restrict ourselves the column labeled by z_j. Suppose that in this column, one entry is twice as large as another; that means that conditioned on $Z = z_j$, we want the probability that Y takes the first value to be twice as large as Y taking the second value. By this reasoning, we see that the conditional probability of $Y = y_i$ conditioned on $Z = z_j$ should be proportional to $\Pr(Y = y_i, Z = z_j)$. And to figure out what this proportion should be, recall that all our probabilities should add up to 1. In other words, we want to divide by $\sum_{i=1}^{n} \Pr(Y = y_i, Z = z_j)$, which is just the sum of the numbers in this column. But notice that if we add up all the numbers in this column, we are just computing $\Pr(Z = z_j)$, by the above observation. All this leads to the following definition:

Definition 1.4. For two random variables Y, Z and two values $y \in \mathcal{Y}, z \in \mathcal{Z}$, the *conditional probability* of $Y = y$ given $Z = z$ is

$$\Pr(Y = y \mid Z = z) = \frac{\Pr(Y = y \text{ and } Z = z)}{\Pr(Z = z)}.$$

Example 1.5. As before, let X be the outcome of a fair die roll, and Y and Z be the indicator variables for the events that X is prime and X is even, respectively. Let's compute $\Pr(Y = 1 \mid Z = 0)$. By the definition,

$$\begin{aligned}
\Pr(Y = 1 \mid Z = 0) &= \frac{\Pr(Y = 1, Z = 0)}{\Pr(Z = 0)} \\
&= \frac{\Pr(X \text{ is prime and not even})}{1/2} \\
&= \frac{\Pr(X = 3 \text{ or } X = 5)}{1/2} \\
&= \frac{2/6}{1/2} = \frac{2}{3}.
\end{aligned}$$

We should also check that this matches our intuition: conditioning on $Z = 0$ is the same as conditioning on X being odd, which means that $X \in \{1, 3, 5\}$. Two of these three numbers are prime, so we indeed expect the conditional probability of $Y = 1$ to be $2/3$.

Observe that if Y and Z are two *independent* random variables, then we have that for any $y \in \mathcal{Y}, z \in \mathcal{Z}$,

$$\Pr(Y = y \mid Z = z) = \frac{\Pr(Y = y, Z = z)}{\Pr(Z = z)} = \frac{\Pr(Y = y)\Pr(Z = z)}{\Pr(Z = z)} = \Pr(Y = y).$$

This is the behavior we expect from independent random variables, for if Y is independent of Z, then conditioning on the value of Z should not affect the value of Y.

1.2 Operations on random variables

We can treat random variables like we do other variables: we can add them, multiply them, and so on. To do any of these operations, just imagine that we use whatever randomness we have to obtain a value for each of our random variables, and then we can apply the desired operation to these values. Since the outcome depended on the outcome of the random choices that were made, the outcome is a random variable.

Example 1.6. As above, let X be the outcome of a fair die roll, and let Y be the indicator variable for the event that X is prime. Let's compute $X+Y$ and XY.

First, to compute $X+Y$, we can simply split into cases depending on what the outcome of the die roll is. If the die comes up 1, then $X=1$ and $Y=0$, so $X+Y=1$. If it comes up 2, then $X=2, Y=1$, so $X+Y=3$. If it comes up 3, then $X=3, Y=1$, so $X+Y=4$. If it comes up 4, then $X=4, Y=0$, so $X+Y=4$ again. If it comes up 5, then $X=5, Y=1$, so $X+Y=6$. Finally, if it comes up 6, then $X=6, Y=0$, so $X+Y=6$ again. Putting this all together, we see that

$$\Pr(X+Y=m) = \begin{cases} \frac{1}{6} & \text{for } m=1, \\ \frac{1}{6} & \text{for } m=3, \\ \frac{1}{3} & \text{for } m=4, \\ \frac{1}{3} & \text{for } m=6. \end{cases}$$

For computing XY, let's try a slightly different approach. Since X can take on any value in $\{1,\dots,6\}$, and since Y can take on the values 0 or 1, we see that XY could in principle take on any value between 0 and 6. For each such value, we'll think of all the ways this value can arise as an outcome of XY, and use this to compute its probability. First, for the outcome 0, this will only arise if $Y=0$ (since X is always positive, and the only way to multiply a positive number by something and get 0 is if the second thing is 0). As we saw above, $\Pr(Y=0)=\frac{1}{2}$, so $\Pr(XY=0)=\frac{1}{2}$ as well. By similar reasoning, we see that each of the potential outcomes $1,4,6$ actually have probability 0: these can never arise as an outcome of XY, since Y must be 0 if X takes on a value in $\{1,4,6\}$. On the other hand, each of the outcomes $2,3,5$ arises with probability $1/6$, since that is the probability that X takes on such a value, in which case $Y=1$. In all,

$$\Pr(XY=m) = \begin{cases} \frac{1}{2} & \text{for } m=0, \\ \frac{1}{6} & \text{for } m=2, \\ \frac{1}{6} & \text{for } m=3, \\ \frac{1}{6} & \text{for } m=5. \end{cases}$$

2 Expectation

Definition 2.1. The *expectation* (also called *mean, expected value,* or *average*) of a random variable X is

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \Pr(X = x).$$

This is a weighted average of the elements of $\mathcal{X}$, where the weights are given by the distribution of X.

Example 2.2. For X the outcome of a fair die roll, we have that

$$\mathbb{E}[X] = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = \frac{7}{2} = 3.5.$$

This suggests that the name "expected value" is somewhat misleading, since when we roll a fair die, we never expect to see the outcome 3.5. However, a deep and important theorem (the so-called *law of large numbers*) says that if we roll a fair die many times and take the average of the outcomes, it will be close to the expectation.

One of the most important properties of the expectation is its *linearity*.

Proposition 2.3. *For any two random variables X, Y, and for any number $\alpha \in \mathbb{R}$, we have*

$$\mathbb{E}[\alpha X] = \alpha \mathbb{E}[X] \qquad \textit{and} \qquad \mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

Proof. For the first equation, we have

$$\mathbb{E}[\alpha X] = \sum_{x \in \mathcal{X}} (\alpha x) \Pr(X = x) = \alpha \sum_{x \in \mathcal{X}} \Pr(X = x) = \alpha \mathbb{E}[X].$$

For the second, we have

$$\begin{aligned}
\mathbb{E}[X + Y] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x + y) \Pr(X = x, Y = y) \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} x \Pr(X = x, Y = y) + y \Pr(X = x, Y = y) \\
&= \sum_{x \in X} x \sum_{y \in \mathcal{Y}} \Pr(X = x, Y = y) + \sum_{y \in \mathcal{Y}} y \sum_{x \in \mathcal{X}} \Pr(X = x, Y = y) \\
&= \sum_{x \in \mathcal{X}} x \Pr(X = x) + \sum_{y \in \mathcal{Y}} y \Pr(Y = y) \\
&= \mathbb{E}[X] + \mathbb{E}[Y].
\end{aligned}$$

In the penultimate equality, we used the property of the table we discussed above: when we add up $\Pr(X = x, Y = y)$ over all possible values of y, we simply recover $\Pr(X = x)$, and similarly the sum of $\Pr(X = x, Y = y)$ over all possible values of x is simply $\Pr(Y = y)$. □

One very important thing to note about this proof is that we did not assume that X and Y were independent, and indeed, the proof works regardless of the dependency between X and Y. I find this pretty counter-intuitive; for instance, suppose X represents the amount of rain on Sunday, and Y represents the amount of rain on Monday. Then these two random variables are dependent (e.g. if a storm is approaching, then we expect them both to be large), but nevertheless the expected total amount of rain over the weekend is just the sum of the expected amount on Sunday and on Monday.

Example 2.4. As above, let X be the outcome of a fair die roll and Y be the indicator variable that X is prime. We computed above that $\mathbb{E}[X] = 3.5$, and we can see that

$$\mathbb{E}[Y] = 0 \cdot \Pr(Y = 0) + 1 \cdot \Pr(Y = 1) = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}.$$

Therefore, $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] = 4$. We can also compute $\mathbb{E}[X + Y]$ directly from the definition, using our computations in Example 1.6. Namely,

$$\mathbb{E}[X + Y] = 1 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{3} + 6 \cdot \frac{1}{3} = 4.$$

We can also compute $\mathbb{E}[XY]$ from the definition and Example 1.6, as

$$\mathbb{E}[XY] = 0 \cdot \frac{1}{2} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} = \frac{5}{3}.$$

Note that $\mathbb{E}[X]\mathbb{E}[Y] = \frac{7}{2} \cdot \frac{1}{2} = \frac{7}{4} \neq \frac{5}{3}$, so we conclude that in general, the expectation of a product is not the product of the expectations. However, the following result shows that for *independent* random variables, this is true.

Proposition 2.5. *If X and Y are independent random variables, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.*

Proof. We can compute directly from the definition that

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (xy) \Pr(X = x \text{ and } Y = y) \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} xy \Pr(X = x) \Pr(Y = y) \\
&= \sum_{x \in \mathcal{X}} x \Pr(X = x) \sum_{y \in \mathcal{Y}} y \Pr(Y = y) \\
&= \sum_{x \in \mathcal{X}} x \Pr(X = x) \mathbb{E}[Y] \\
&= \mathbb{E}[Y] \sum_{x \in \mathcal{X}} x \Pr(X = x) \\
&= \mathbb{E}[Y]\mathbb{E}[X].
\end{aligned}$$

□

Another important notion for us will be that of *conditional expectation*. We need to define it in two stages. First, we define $\mathbb{E}[X \mid Y = y]$:

Definition 2.6. For two random variables X, Y, and some value $y \in \mathcal{Y}$, we define

$$\mathbb{E}[X \mid Y = y] = \sum_{x \in \mathcal{X}} x \Pr(X = x \mid Y = y)$$

Equivalently, observe that we can think of $(X \mid Y = y)$ as a random variable: it takes on the value $x \in \mathcal{X}$ with probability $\Pr(X = x \mid Y = y)$. In this case, $\mathbb{E}[X \mid Y = y]$ is just the ordinary expectation of this random variable.

Definition 2.7. For two random variables X, Y, their conditional expectation $\mathbb{E}[X \mid Y]$ is defined as the random variable that takes on the value $\mathbb{E}[X \mid Y = y]$ with probability $\Pr(Y = y)$, for all $y \in \mathcal{Y}$.

Note that up to now, all our expectations have been real numbers, whereas this new quantity $\mathbb{E}[X \mid Y]$ is itself another random variable. One way to think about it is to imagine randomly determining a value y for Y, and then having the new random variable take the value $\mathbb{E}[X \mid Y = y]$.

How should we think about this random variable? One way is as follows. Note that $\mathbb{E}[X \mid Y = y]$ means something fairly intuitive: it is the average we expect for X, given that we know that Y took on the value y. We can think of this as our "best guess" for X, once we know the information $Y = y$. Now, $\mathbb{E}[X \mid Y]$ is a random variable that records what our best guess for X *would be*, if someone told us the value of Y. But of course, no one has told us the value of Y, so this is a random quantity: there is some probability that Y takes on the value y, and in this case our best guess would be $\mathbb{E}[X \mid Y = y]$.

Example 2.8. As before, let X be the outcome of a fair die roll, let Y be the indicator random variable that X is prime, and let Z be the indicator random variable that X is even. For $x \in \{1, \ldots, 6\}$, note that $\Pr(X = x \mid Y = 0)$ is equal to 0 if x is prime, and $1/3$ otherwise. Therefore,

$$\mathbb{E}[X \mid Y = 0] = \sum_{x \in \mathcal{X}} x \Pr(X = x \mid Y = 0) = 1 \cdot \frac{1}{3} + 4 \cdot \frac{1}{3} + 6 \cdot \frac{1}{3} = \frac{11}{3}.$$

Similarly, we can compute that $\mathbb{E}[X \mid Y = 1] = 2 \cdot \frac{1}{3} + 3 \cdot \frac{1}{3} + 5 \cdot \frac{1}{3} = \frac{10}{3}$. Therefore, $\mathbb{E}[X \mid Y]$ is a random variable that takes on the value $11/3$ with probability $\frac{1}{2}$, and the value $10/3$ with probability $\frac{1}{2}$.

With a similar computation, we can see that $\mathbb{E}[X \mid Z]$ takes on the value 3 with probability $\frac{1}{2}$ and the value 4 with probability $\frac{1}{2}$. Additionally, $\mathbb{E}[Y \mid Z]$ takes on the values $1/3$ and $2/3$, each with probability $\frac{1}{2}$.

Finally, we can calculate that $\mathbb{E}[Y \mid X]$ takes on the values 0 and 1, each with probability $\frac{1}{2}$. Indeed, the reason for this is that if we condition on the event that $X = x$, then the value of Y is completely determined—it's 1 if x is prime, and 0 if not. So conditioning on X does nothing, and we have that $\mathbb{E}[Y \mid X] = Y$. For the same reason, $\mathbb{E}[Z \mid X] = Z$.

The last example shows that if some random variable Y is *determined* by a random variable X (i.e. if knowing the value of X completely eliminates any randomness from the value of Y), then $\mathbb{E}[Y \mid X] = Y$. In the opposite extreme, if X and Y are independent, then $\mathbb{E}[Y \mid X] = \mathbb{E}[Y]$. Note that while $\mathbb{E}[Y \mid X]$ is in general a random variable, in this case, it is simply a (non-random) number, namely the number $\mathbb{E}[Y]$. The following theorem includes these two facts, as well as a few other important properties we will need of conditional expectation.

Theorem 2.9. *Let X, Y, Z be random variables.*

(i) Conditional expectation is linear: $\mathbb{E}[X + Y \mid Z] = \mathbb{E}[X \mid Z] + \mathbb{E}[Y \mid Z]$ and $\mathbb{E}[\alpha X \mid Z] = \alpha\mathbb{E}[X \mid Z]$ for any real number α.

(ii) If X is determined by Z, then $\mathbb{E}[X \mid Z] = X$.

(iii) If X and Z are independent, then $\mathbb{E}[X \mid Z] = \mathbb{E}[X]$.

(iv) Tower property: $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Z]]$.

(v) Taking out what is known: if Y is determined by Z, then $\mathbb{E}[XY \mid Z] = Y\mathbb{E}[X \mid Z]$.

Note that Theorem 2.9(ii) is a special case of Theorem 2.9(v). Perhaps the trickiest property to internally absorb is the tower property. Recall that $\mathbb{E}[X \mid Z]$ is a random variable; then Theorem 2.9(iv) says that X and $\mathbb{E}[X \mid Z]$ are both random variables with the same expected value.

Proof. Parts (i)–(iii) are left as exercises, and we only prove the final two parts.

(iv) Recall that $\mathbb{E}[X \mid Z]$ takes on the value $\mathbb{E}[X \mid Z = z]$ with probability $\Pr(Z = z)$, for all $z \in \mathcal{Z}$. Therefore,

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[X \mid Z]] &= \sum_{z \in \mathcal{Z}} \mathbb{E}[X \mid Z = z] \Pr(Z = z) \\
&= \sum_{z \in \mathcal{Z}} \left(\sum_{x \in \mathcal{X}} x \Pr(X = x \mid Z = z) \right) \Pr(Z = z) \\
&= \sum_{x \in \mathcal{X}} x \sum_{z \in \mathcal{Z}} \Pr(Z = z) \Pr(X = x \mid Z = z) \\
&= \sum_{x \in \mathcal{X}} x \sum_{z \in \mathcal{Z}} \Pr(X = x \text{ and } Z = z) \\
&= \sum_{x \in \mathcal{X}} x \Pr(X = x) \\
&= \mathbb{E}[X],
\end{aligned}$$

where we use the definition of conditional probability in the fourth line, and the column-sum property we observed earlier in the fifth line.

(v) By definition, the random variable $\mathbb{E}[XY \mid Z]$ takes on the value $\mathbb{E}[XY \mid Z = z]$ with probability $\Pr(Z = z)$, for all $z \in \mathcal{Z}$. Fix some $z \in \mathcal{Z}$. Then by definition,

$$\mathbb{E}[XY \mid Z = z] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} xy \Pr(X = x \text{ and } Y = y \mid Z = z).$$

Since Y is determined by Z, if we know that $Z = z$, then there is no randomness left in deciding whether $Y = y$ or $Y \neq y$. In other words, $\Pr(Y = y \mid Z = z)$ equals either zero or one. As such, we observe that

$$\Pr(X = x \text{ and } Y = y \mid Z = z) = \Pr(X = x \mid Z = z) \Pr(Y = y \mid Z = z). \qquad (*)$$

Indeed, if $\Pr(Y = y \mid Z = z) = 0$ (i.e. if $Y = y$ definitely *does not* happen given $Z = z$), then both sides of $(*)$ are 0, so this is certainly true. On the other hand, if $\Pr(Y = y \mid Z = z) = 1$ (i.e. if $Y = y$ definitely *does* happen given $Z = z$), then both sides of $(*)$ equal $\Pr(X = x \mid Z = z)$, so it's again true.

Plugging in $(*)$ into the definition of $\mathbb{E}[XY \mid Z = z]$, we find that

$$\begin{aligned}
\mathbb{E}[XY \mid Z = z] &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} xy \Pr(X = x \mid Z = z) \Pr(Y = y \mid Z = z) \\
&= \left(\sum_{y \in \mathcal{Y}} y \Pr(Y = y \mid Z = z) \right) \left(\sum_{x \in \mathcal{X}} x \Pr(X = x \mid Z = z) \right) \\
&= \mathbb{E}[Y \mid Z = z] \mathbb{E}[X \mid Z = z].
\end{aligned}$$

So in other words, the random variable $\mathbb{E}[XY \mid Z]$ takes on the value $\mathbb{E}[Y \mid Z = z]\mathbb{E}[X \mid Z = z]$ with probability $\Pr(Z = z)$. The final observation is that by Theorem 2.9(ii), the random variable which equals $\mathbb{E}[Y \mid Z = z]$ with probability $\Pr(Z = z)$ is simply Y, so we can conclude that $\mathbb{E}[XY \mid Z] = Y\mathbb{E}[X \mid Z]$. □

3 Martingales

With all this setup, we are finally ready to define and study martingales.

Definition 3.1. Let $X_0, X_1, X_2, \ldots$ and $Y_0, Y_1, Y_2, \ldots$ be two (finite or infinite) sequences of random variables. We say that $X_0, X_1, \ldots$ is a *martingale with respect to* $Y_0, Y_1, \ldots$ if the following two properties hold. First, for all $i \geq 0$, the random variable X_i is determined by $Y_0, \ldots, Y_i$, and second, for all $i \geq 1$,

$$\mathbb{E}[X_i \mid Y_0, Y_1, \ldots, Y_{i-1}] = X_{i-1}.$$

How should we think about a martingale? Intuitively, it is supposed to model a fair game. In this model, the random variables $Y_0, Y_1, \ldots$ are the outcomes of the randomness in the game (e.g. the spins of a roulette wheel, or the outcomes of coin flips, or the draws from

a shuffled deck of cards). Additionally, X_i represents your total earnings (or losses) after the ith round of the game. Then the martingale condition says that the ith round is fair: on average, you will neither make nor lose money in the ith round, so that your average total winnings after the ith round, conditioned on everything that happened before, equal your current total winnings.

Example 3.2. Let $(Y_i)_{i\geq 0}$ be a sequence of independent random variables, with

$$\Pr(Y_i = 1) = \Pr(Y_i = -1) = \frac{1}{2}.$$

Then Y_i models the game in which at every turn we flip a fair coin, and you win \$1 if it comes up heads and lose \$1 if it comes up tails. Let $X_i = Y_0 + Y_1 + \cdots + Y_i$, which denotes your total winnings after the ith round of the game. Then we claim that $(X_i)_{i\geq 0}$ is a martingale with respect to $(Y_i)_{i\geq 0}$. The first condition is immediate, since X_i is defined in terms of $Y_0, \ldots, Y_i$, so it is certainly determined by them. Additionally, by Theorems 2.9(i), 2.9(iii) and 2.9(v), we have that

$$\begin{aligned}
\mathbb{E}[X_i \mid Y_0, \ldots, Y_{i-1}] &= \mathbb{E}[Y_0 + \cdots + Y_i \mid Y_0, \ldots, Y_{i-1}] \\
&= \mathbb{E}[Y_0 \mid Y_0, \ldots, Y_{i-1}] + \mathbb{E}[Y_1 \mid Y_0, \ldots, Y_{i-1}] + \cdots + \mathbb{E}[Y_i \mid Y_0, \ldots, Y_{i-1}] \\
&= Y_0 + Y_1 + \cdots + Y_{i-1} + \mathbb{E}[Y_i] \\
&= Y_0 + Y_1 + \cdots + Y_{i-1} \\
&= X_{i-1},
\end{aligned}$$

using the fact that $Y_0, \ldots, Y_{i-1}$ are determined by themselves, that Y_i is independent of $Y_0, \ldots, Y_{i-1}$, and that $\mathbb{E}[Y_i] = 0$.

Note that in this example, we never actually used that each Y_i takes on the values ± 1 with probability $1/2$: the only properties we used were independence and the fact that $\mathbb{E}[Y_i] = 0$. So the same thing works in general: if $Y_0, Y_1, \ldots$ are independent random variables each of which has expectation zero, then their partial sums form a martingale. This should be thought of as modeling *any* fair game.

Example 3.3. Let $(Y_i)_{i\geq 0}$ be independent random variables, with $\mathbb{E}[Y_i] = 1$ for all i. Let $X_i = Y_0 Y_1 \cdots Y_i$. Then we claim that $(X_i)_{i\geq 0}$ is a martingale with respect to $(Y_i)_{i\geq 0}$. The determinedness condition is again immediate. To check the martingale condition, we note that

$$\begin{aligned}\mathbb{E}[X_i \mid Y_0, \ldots, Y_{i-1}] &= \mathbb{E}[Y_0 Y_1 \cdots Y_{i-1} Y_i \mid Y_0, \ldots, Y_{i-1}] \\ &= Y_0 Y_1 \cdots Y_{i-1} \mathbb{E}[Y_i \mid Y_0, \ldots, Y_{i-1}] \\ &= Y_0 Y_1 \cdots Y_{i-1} \mathbb{E}[Y_i] \\ &= Y_0 Y_1 \cdots Y_{i-1} \\ &= X_{i-1},\end{aligned}$$

where the second line uses Theorem 2.9(v), the third uses the independence of Y_i from $Y_0, \ldots, Y_{i-1}$, and the fourth uses our assumption that $\mathbb{E}[Y_i] = 1$.

This example can be thought of as a "multiplicative" version of the previous fair game example. In this case, your total earnings get multiplied by some random quantity at every step (e.g. the stock market might go up or down by some percentage, and thus your investments get multiplied by some factor), and our assumption is that these steps are independent, and that the average multiplier is 1. Then we again have the martingale condition.

Most of the results we will prove about martingales concern our intuition that "you can't beat the system". If you are playing a fair game, then there is no strategy that gets you net winnings on average. We will prove increasingly robust versions of this idea, culminating in Doob's optional stopping theorem. But we begin with the following simple observation, which says that your average worth after n steps of a fair game is equal to your worth when you started.

Theorem 3.4. *Let $X_0, X_1, \ldots$ be a martingale with respect to some sequence $Y_0, Y_1, \ldots$. Then for any n,*

$$\mathbb{E}[X_n] = \mathbb{E}[X_0].$$

Proof. We prove this by induction on n, with the base case $n = 0$ being immediate since it just says $\mathbb{E}[X_0] = \mathbb{E}[X_0]$. Now suppose we have proved that $\mathbb{E}[X_{n-1}] = \mathbb{E}[X_0]$, and we wish to prove the same for $\mathbb{E}[X_n]$. Using the tower property of conditional expectation (Theorem 2.9(iv)), we have that

$$\mathbb{E}[X_n] = \mathbb{E}[\mathbb{E}[X_n \mid Y_0, \ldots, Y_{n-1}]] = \mathbb{E}[X_{n-1}] = \mathbb{E}[X_0],$$

where the second equality is the martingale condition, and the third is our inductive hypothesis. □

4 The martingale transform

In this section, we discuss one of the most important operations one can do on martingales, called the *martingale transform*. For intuition, recall the basic coin-flipping game we started

with: at every step, we flip a fair coin, and we gain \$1 if it comes up heads and lose \$1 if it comes up tails. In a real casino, we'd have the option to choose how much we wager: maybe we want to bet \$1000 on a certain coin toss, but we want to sit the next one out (and thus wager \$0 on it).

Suppose that $Y_0, Y_1, \dots$ are the outcomes of the independent coin tosses, i.e.

$$\Pr(Y_i = 1) = \Pr(Y_i = -1) = \frac{1}{2}.$$

If we wager w_i dollars on the ith coin toss, then our winnings after the nth step are $w_0Y_0 + w_1Y_1 + \cdots + w_nY_n$. It is not hard to check that the sequence of random variables defined by $X_n = w_0Y_0 + \cdots + w_nY_n$ forms a martingale, for any real numbers $w_0, w_1, \dots$.

However, in real life, we don't actually commit to the wagers *before* the game starts. Instead, we probably want to choose our bets based on what's already happened in the game. For example, many people who bet on roulette watch to see a long streak of red outcomes (for example), and then bet on black, using the assumption that "black is due for a win". So we should allow the wagers themselves to be random variables, which are allowed to depend on the randomness in the sequence $Y_0, Y_1, \dots$.

However, we don't want to allow the wagers to depend *arbitrarily* on the sequence of outcomes. For example, it'd be great to have a betting strategy of the form "bet \$1000 on a flip if it'll come up heads, and otherwise bet \$0 on it". This betting strategy does *depend* on the outcomes of the coin flips, but it requires looking into the future! So the wager at the nth step should only be allowed to depend on the past, i.e. on the outcomes of the coin tosses that happened *before* time n. This motivates the following definition.

Definition 4.1. A sequence $W_1, W_2, \dots$ is called *prophecy-free* with respect to another sequence $Y_0, Y_1, \dots$ if for all $n \geq 1$, the random variable W_n is determined by the random variables $Y_0, \dots, Y_{n-1}$.

We call these sequences prophecy-free because they don't require looking into the future.

Definition 4.2. Let $(Y_i)_{i\geq 0}$ be a sequence of random variables. Let $(X_i)_{i\geq 0}$ be a martingale with respect to $(Y_i)_{i\geq 0}$, and let $(W_i)_{i\geq 1}$ be a prophecy-free with respect to $(Y_i)_{i\geq 0}$. Then the *martingale transform* of X by W is the sequence $((W \bullet X)_i)_{i\geq 0} = (W \bullet X)_0, (W \bullet X)_1, \dots$ defined inductively by

$$(W \bullet X)_0 = X_0; \qquad (W \bullet X)_i = (W \bullet X)_{i-1} + W_i(X_i - X_{i-1}).$$

Unwrapping this definition, it says that for $i \geq 1$,

$$(W \bullet X)_i = X_0 + \sum_{k=1}^{i} W_k(X_k - X_{k-1}).$$

The definition of the martingale transform can be a bit overwhelming at first, but it really is the same definition we had above. On the ith round of a game, your total money changes

by $X_i - X_{i-1}$. If you had wagered W_i dollars on the ith round, then your winnings would instead change by $W_i(X_i - X_{i-1})$. Adding this up over all rounds, we see that $(W \bullet X)_i$ is exactly your total money after the ith round of the game, where your wagering is given by the sequence $W_1, W_2, \ldots$.

Our main theorem in this section is that as long as your wagering sequence is prophecy-free, you still can't beat the system: no matter how you wager (without cheating by looking into the future) your average net winnings will still be zero.

Theorem 4.3. *Let $(X_i)_{i\geq 0}$ be a martingale and $(W_i)_{i\geq 1}$ be a prophecy-free sequence, both respect to $(Y_i)_{i\geq 0}$. Then the transformed sequence $((W\bullet X)_i)_{i\geq 0}$ also a martingale with respect to $(Y_i)_{i\geq 0}$.*

Proof. We begin by checking that $(W \bullet X)_i$ is determined by $Y_0, \ldots, Y_i$, for all $i \geq 0$. Indeed, this is clear for $i = 0$, since $(W \bullet X)_0 = X_0$ and X_0 is determined by Y_0 from our assumption that $(X_i)_{i\geq 0}$ is a martingale. For $i \geq 1$, we have that

$$(W \bullet X)_i = X_0 + \sum_{k=1}^{i} W_k(X_k - X_{k-1}).$$

Every variable appearing in this expression is either of the form W_k for $k \leq i$ or X_k for $k \leq i$. By assumption, each of these is determined by $Y_0, \ldots, Y_i$, which proves the first condition.

For the martingale condition, we can compute that for $i \geq 1$,

$$\begin{aligned}\mathbb{E}[(W \bullet X)_i \mid Y_0, \ldots, Y_{i-1}] &= \mathbb{E}[(W \bullet X)_{i-1} + W_i(X_i - X_{i-1}) \mid Y_0, \ldots, Y_{i-1}] \\ &= \mathbb{E}[(W \bullet X)_{i-1} \mid Y_0, \ldots, Y_{i-1}] + \mathbb{E}[W_i(X_i - X_{i-1}) \mid Y_0, \ldots, Y_{i-1}].\end{aligned}$$

By our argument above, we know that $(W \bullet X)_{i-1}$ is determined by $Y_0, \ldots, Y_{i-1}$. So by Theorem 2.9(ii), we have that the first term in the sum above is simply $(W \bullet X)_{i-1}$. For the second term, we recall that W_i is determined by $Y_0, \ldots, Y_{i-1}$, by the definition of prophecy-free. So by Theorem 2.9(v),

$$\mathbb{E}[W_i(X_i - X_{i-1}) \mid Y_0, \ldots, Y_{i-1}] = W_i\mathbb{E}[(X_i - X_{i-1}) \mid Y_0, \ldots, Y_{i-1}].$$

Additionally, we see that

$$\mathbb{E}[(X_i - X_{i-1}) \mid Y_0, \ldots, Y_{i-1}] = \mathbb{E}[X_i \mid Y_0, \ldots, Y_{i-1}] - \mathbb{E}[X_{i-1} \mid Y_0, \ldots, Y_{i-1}] = X_{i-1} - X_{i-1} = 0.$$

Putting this all together, we find that

$$\mathbb{E}[(W \bullet X)_i \mid Y_0, \ldots, Y_{i-1}] = (W \bullet X)_{i-1},$$

proving the martingale property. □

As a simple corollary, we can see that no prophecy-free betting system can win you money on average.

Corollary 4.4 (No prophet, no profit)**.** *Let $(X_i)_{i\geq 0}$ be a martingale and $(W_i)_{i\geq 1}$ be a prophecy-free sequence, both respect to $(Y_i)_{i\geq 0}$. Then for any $n \geq 0$,*

$$\mathbb{E}[(W \bullet X)_n] = \mathbb{E}[X_0].$$

Proof. We know that $((W \bullet X)_i)_{i\geq 0}$ is a martingale, and that $(W \bullet X)_0 = X_0$ by definition. So by Theorem 3.4,

$$\mathbb{E}[(W \bullet X)_n] = \mathbb{E}[(W \bullet X)_0] = \mathbb{E}[X_0]. \qquad \square$$

5 Stopping times

We saw in the last section that using a betting strategy that varies your wagers can never let you beat the system in a fair game. However, there is another type of strategy you can use, which is to simply leave the game at a good time, in the hopes that we pick this time appropriately in order to maximize our earnings. Stopping strategies we might use include "the first time we have \$100" or "after winning 5 rounds in a row" or "after the third loss". However, just as before, we shouldn't be allowed to predict the future, so we can't use stopping rules like "when our winnings are as large as they will ever be" or "the last time we have at least \$1000"—such stopping rules require knowing the outcomes of future rounds, and of course we can't use that in deciding when to quit. This motivates the following definition, similarly to our earlier definition of prophecy-free sequences.

Definition 5.1. Let $(Y_i)_{i\geq 0}$ be a sequence of random variables. A *stopping time* with respect to $(Y_i)_{i\geq 0}$ is a random variable T taking values in $\{0, 1, 2, \dots\} \cup \{\infty\}$, with the following property. For every $n \geq 0$, the variables $Y_0, \dots, Y_n$ determine whether $T \leq n$ or $T > n$.

In other words, the decision of whether we leave the casino before round $n + 1$—namely the event that $T \leq n$—depends only on the outcomes of the first n rounds, and not on the future. Note that we allow T to take on the value ∞, which simply corresponds to never leaving the casino.

Definition 5.2. Let $(X_i)_{i\geq 0}$ be a martingale and let T be a stopping time, both with respect to some sequence $(Y_i)_{i\geq 0}$. The *stopped process* $(X_i^T)_{i\geq 0}$ is defined by

$$X_i^T = X_{\min(i,T)} = \begin{cases} X_i & \text{if } i \leq T, \\ X_T & \text{if } i > T. \end{cases}$$

In other words, the stopped process follows the original process until the stopping time happens. At that point, we leave the casino, and the value of our sequence no longer changes. As a simple example, suppose that T is the constant random variable which always takes on the value 4; in other words, we play through the fourth round of the game, and then leave the casino. Then the stopped process is simply $X_0, X_1, X_2, X_3, X_4, X_4, X_4, \dots$: our winnings after the fourth round remain our winnings forever.

The next result says that the stopped process is another martingale. It *almost* says that a stopping strategy can't let you beat a fair game, but not quite; we'll soon see what this "almost" entails.

Theorem 5.3. *Let $(X_i)_{i\geq 0}$ be a martingale and let T be a stopping time, both with respect to some sequence $(Y_i)_{i\geq 0}$. The stopped process $(X_i^T)_{i\geq 0}$ is also a martingale with respect to $(Y_i)_{i\geq 0}$.*

Proof. For $i \geq 1$, let's define W_i to be the indicator random variable of whether $T \geq i$: if $T \geq i$ then $W_i = 1$, and otherwise $W_i = 0$. The value of W_i is determined by whether $T \geq i$, or equivalently it's determined by whether $T \leq i-1$. This is in turn determined by $Y_0, \ldots, Y_{i-1}$, by the definition of a stopping time. This shows that W_i is determined by $Y_0, \ldots, Y_{i-1}$, so $(W_i)_{i\geq 1}$ is a prophecy-free sequence. Therefore, by Theorem 4.3, the sequence $((W \bullet X)_i)_{i\geq 0}$ is a martingale. By definition of the martingale transform, we have that $(W \bullet X)_0 = X_0 = X_0^T$, and that for $i \geq 1$,

$$(W \bullet X)_i = X_0 + \sum_{k=1}^{i} W_k(X_k - X_{k-1}) = X_0 + W_1(X_1 - X_0) + W_2(X_2 - X_1) + \cdots + W_i(X_i - X_{i-1}).$$

W_k equals 1 as long as $k \leq T$, and afterwards $W_k = 0$. If $i \leq T$, then all the W_k above equal 1, so the right-hand side simplifies to

$$X_0 + (X_1 - X_0) + (X_2 - X_1) + \cdots + (X_i - X_{i-1}) = X_i.$$

On the other hand, if $i > T$, then the right-hand side simplifies to

$$X_0 + (X_1 - X_0) + (X_2 - X_1) + \cdots + (X_T - X_{T-1}) = X_T.$$

In either case, we see that $(W \bullet X)_i = X_i^T$, i.e. the stopped process simply equals the martingale transform by the prophecy-free sequence $(W_i)_{i\geq 1}$. By Theorem 4.3, we conclude that the stopped process $(X_i^T)_{i\geq 0}$ is a martingale. □

Corollary 5.4. *Let $(X_i)_{i\geq 0}$ be a martingale and let T be a stopping time, both with respect to some sequence $(Y_i)_{i\geq 0}$. For any $n \geq 0$, we have that*

$$\mathbb{E}[X_n^T] = \mathbb{E}[X_{\min(n,T)}] = \mathbb{E}[X_0].$$

Proof. The first equality is simply the definition of X_n^T. For the second, we have that $(X_i^T)_{i\geq 0}$ is a martingale, and that $X_0^T = X_0$. So we get the second equality by Theorem 3.4. □

This seems to say that you can't win, on average, by stopping your play at some point. And indeed it does roughly say that: the value X_n^T represents your winnings at time n, assuming you leave the casino whenever the stopping rule T takes effect. However, this isn't the actual quantity we care about. What we care about is the random variable X_T: this represents your winnings when you leave the casino. The following three examples show that this subtle distinction is actually very important.

Example 5.5. Let $Y_0 = 0$, and let $(Y_i)_{i\geq 1}$ be a sequence of independent random variables, each taking the value ± 1 with probability $1/2$. Note that we artifically make $Y_0 = 0$, so that Y_i represents the ith round of the game.

1. Let $X_i = Y_0 + \cdots + Y_i$ be our winnings if we play the usual game, either winning or losing \$1 at every round. Let T be the first time i for which $|X_i| = 2$. In other words, T is the first time where our net earnings or losses are \$2. It is easy to see that T is a stopping time, so the stopped process $(X_i^T)_{i\geq 1}$ is a martingale. Then Corollary 5.4 shows that $\mathbb{E}[X_n^T] = \mathbb{E}[X_0] = 0$, meaning that after the ith round, our average winnings are zero, regardless of whether or not we left the casino by that point.

   Additionally, by symmetry, it is certainly believable that when we leave the casino, we are equally likely to have gained \$2 as to have lost \$2. This means that X_T takes on the values ± 2 with equal probability, implying that $\mathbb{E}[X_T] = 0$. Thus, in this case, we really haven't beat the system: on average we win nothing when we leave the casino.

2. Let $(X_i)_{i\geq 0}$ be as in the last example, but in this case let S be the first time i that $X_i = 2$. In other words, S represents the stopping strategy where we keep betting until our total earnings equal \$2, and then we leave. This is again a stopping time, so we have that $\mathbb{E}[X_n^S] = \mathbb{E}[X_0] = 0$ by Corollary 5.4. However, it is also clear that X_S equals 2 with probability 1: under this stopping rule, when we leave the casino, we definitely do so with \$2 in hand (of course, it's possible that we never leave). So in this case, $\mathbb{E}[X_S] = 2$, and thus we have made money!

3. Finally, let
$$Z_i = 2Y_1 + 4Y_2 + \cdots + 2^i Y_i.$$
   In other words, this represents our winnings in the doubling game, where at every step we bet twice as much money. Let R denote the first time i for which $Y_i = 1$, i.e. the first time that a heads comes up. R is again a stopping time, and this time there are no weird shenanigans where R might never happen: it is intuitively clear (and not too hard to rigorously prove) that the probability of flipping infinitely many tails in a row is zero. However, we claim that Z_R is again the constant 2. Indeed, if i is the first time that a heads comes up (i.e. if $R = i$), then
$$Z_R = Z_i = 2(-1) + 4(-1) + \cdots + 2^{i-1}(-1) + 2^i(1) = -(2^i - 2) + 2^i = 2.$$
   So we again have that $\mathbb{E}[Z_R] = 2$, i.e. that we make money in expectation when we leave the casino, despite Corollary 5.4.

What's going on in these examples? Well, one can prove that in all three cases, the stopping condition will eventually happen with probability 1 (this is very believable for every example but the second, and you'll rigorously prove it for the second on the homework). Therefore, as i tends to infinity, the random variable X_i^T converges to X_T. Indeed, the

sequence $(X_i^T)_{i\geq 0}$ looks like $X_0, X_1, X_2, \ldots, X_{k-1}, X_k, X_k, X_k, \ldots$ for some k depending on the random outcomes, and $T = k$ given these random outcomes. So we have that

$$\lim_{i\to\infty} X_i^T = X_T.$$

It would be great if we could apply expectations to both sides and conclude that

$$\lim_{i\to\infty} \mathbb{E}[X_i^T] = \mathbb{E}[X_T].$$

Unfortunately, this is simply not true in general! Instead, actually applying expectations to both sides above gives

$$\mathbb{E}\left[\lim_{i\to\infty} X_i^T\right] = \mathbb{E}[X_T],$$

and in general, one can't interchange a limit with an expectation. This is an extremely annoying issue, and while it might sound like a stupid technicality, it's actually crucial: Examples 5.5(2) and 5.5(3) above show that such an interchange is simply false in general.

There are several deep and important theorems that give useful conditions for when one *can* perform such an interchange, the most important of which are called the Monotone convergence theorem and the Dominated convergence theorem. We won't state or prove them in this class—doing so requires developing a lot more theory. However, we will state Doob's optional stopping theorem, which is a direct corollary of these. It gives simple, easy-to-check conditions which imply that such an interchange is allowed, and thus that the random variable X_T has the same expectation as X_0.

6 Doob's optional stopping theorem, and applications

Theorem 6.1 (Optional stopping theorem)**.** *Let $(X_i)_{i\geq 0}$ be a martingale and let T be a stopping time, both with respect to some sequence $(Y_i)_{i\geq 0}$. Suppose that (at least) one of the following three conditions holds.*

(a) (T is bounded.) There exists some integer N so that $\Pr(T < N) = 1$.

(b) (T is finite and X is bounded.) There exists some integer K so that $|X_i| \leq K$ for all i, and $\Pr(T = \infty) = 0$.

(c) (T has finite expectation and X has bounded differences.) There exists some integer K so that $|X_i - X_{i-1}| \leq K$ for all i, and $\mathbb{E}[T] < \infty$.

Then $\mathbb{E}[X_T] = \mathbb{E}[X_0]$.

Not really proof. In all three cases, we must have that $\Pr(T = \infty) = 0$, so as discussed above, we know that $\lim_{i\to\infty} X_i^T = X_T$, and thus that $\mathbb{E}[X_T] = \mathbb{E}[\lim_{i\to\infty} X_i^T]$. We would like to interchange the limit and the expectation, but in general we cannot. In the first case, if T is bounded by some integer N, then we actually already know that $X_N^T = X_T$, so we automatically get that $\mathbb{E}[X_T] = \mathbb{E}[X_N^T] = \mathbb{E}[X_0]$ by Corollary 5.4.

In each of the remaining two cases, the Dominated convergence theorem guarantees that

$$\mathbb{E}\left[\lim_{i\to\infty} X_i^T\right] = \lim_{i\to\infty} \mathbb{E}[X_i^T]$$

(i.e. that interchanging the limit and the expectation is OK), which yields the desired result since $\lim_{i\to\infty} X_i^T = X_T$. □

We will now see a number of examples of the optional stopping theorem in action. It is a remarkable result: in a lot of examples, it turns what looks like a hopelessly complicated problem into a two-line computation. Additionally, it is even informative when it *doesn't* hold; we will see some cases where $\mathbb{E}[X_T] \neq \mathbb{E}[X_0]$, from which we can conclude that the none of the assumptions of the optional stopping theorem hold, which in turn tells us something about the process in question.

Example 6.2.

0. Revisiting Example 5.5 In Example 5.5, we had two examples where the value when we leave the casino does not have expectation equal to $\mathbb{E}[X_0]$, so those two examples had better not satisfy the assumptions of the optional stopping theorem. We'll return to example 2 shortly, but let's quickly dispense with example 3, where we keep doubling our wagers until we get a heads. The stopping time is certainly not bounded, so condition (a) doesn't hold. The stopping time is finite with probability 1, but it's certainly not the case that the process is bounded—we might have to go into a great deal of debt before we get heads, so condition (b) also doesn't hold. Finally, while the stopping time has finite expectation (you should check this yourself!), it's also not the case that the differences are bounded: we keep doubling our wager every time, so the differences grow to infinity, and thus condition (c) also fails to hold.

In a certain sense, this explains what goes wrong with with the "always double" betting strategy. Because the optional stopping theorem fails to hold, it must really be the case that the failure modes will come into play. In this case, you'd need to violate either condition (b) or (c). All casinos have a cap on how much you can bet on any game, and thus they enforce the bounded differences condition in (c): you can't make arbitrarily large wagers. Additionally, there is only finitely much money in the world (and you have an even smaller finite amount), so you can't go into arbitrarily much debt, which enforces the boundedness of X in condition (b).

1. The lost frog As in Example 5.5, let $Y_0 = 0$ and Y_i be independent random variables taking on the values ± 1 with probability $1/2$ for all $i \geq 1$, and let $X_i = Y_0 + \cdots + Y_i$. Given two positive integers a, b, let T denote the first time that $X_i \in \{a, -b\}$. This represents the first time where you've either won $\$a$ in total, or lost $\$b$ in total. It is perhaps not too surprising, and you'll prove it on the homework, that $\mathbb{E}[T] < \infty$. This implies that $\mathbb{E}[X_T] = 0$. But X_T takes on either the value a or the value $-b$, with some probabilities p_W and p_L, respectively. So we have that

$$0 = \mathbb{E}[X_T] = p_W \cdot a + p_L \cdot (-b).$$

Plugging in $p_L = 1 - p_W$ and solving as we did at the very beginning of the class, we find what we already found there: the probability of winning $\$a$ before you lose $\$b$ is exactly $b/(a+b)$.

2. Recurrence Let $(Y_i)_{i\geq 0}, (X_i)_{i\geq 0}$ be as in the last example. Note that for any positive integers a, b,

$$\Pr((X_i)_{i\geq 0} \text{ eventually reaches } a) \geq \Pr((X_i)_{i\geq 0} \text{ reaches } a \text{ before reaching } -b) = \frac{b}{a+b}.$$

But this holds for *any* b. Letting b tend to infinity, we conclude that

$$\Pr((X_i)_{i\geq 0} \text{ eventually reaches } a) \geq \lim_{b\to\infty} \frac{b}{a+b} = 1.$$

In other words, with probability 1, every integer is eventually hit by the sequence (X_i). In particular, if we let T be the first time that we reach \$2 (as in Example 5.5(2)), then we see that $\Pr(T = \infty) = 0$. However, we cannot apply condition (b), because the sequence (X_i) is unbounded; we might have to go into a lot of debt before we ever gain \$2.

On the other hand, the sequence (X_i) *does* have bounded differences, so it seems like we might be able to apply condition (c), which would be a contradiction to the fact that $\mathbb{E}[X_T] = 2$. Thus, we can't apply it, and we conclude that we must have $\mathbb{E}[T] = \infty$. In other words, we have the following very weird situation: we will definitely reach \$2 at some point, but on average, we have to wait infinitely long for this to happen!

Note that we can actually iterate the above argument. With probability 1, we get to any fixed value a at some point. But once we get to a, our process looks as though we'd just started at a, so by the same logic, we'll get back to 0 eventually with probability 1. But then once we get back to 0, we'll eventually get back to a with probability 1. We can continue in this way forever, and we conclude that the process $(X_i)_{i\geq 0}$ visits every integer infinitely often with probability 1. This property is known as the *recurrence of the one-dimensional simple random walk*.

3. How long until we stop? Let $(Y_i)_{i\geq 0}, (X_i)_{i\geq 0}$ be as in the last example, and again let T be the first time we reach a or $-b$. Let's compute $\mathbb{E}[T]$, that is, the expected amount of time until we gain $\$a$ or lose $\$b$.

To do this, consider the sequence $(V_i)_{i\geq 0}$ defined by

$$V_i = X_i^2 - i.$$

We claim that $(V_i)_{i\geq 0}$ is a martingale with respect to $(Y_i)_{i\geq 0}$. The fact that V_i is determined by $Y_0, \ldots, Y_i$ is clear, but we need to check the martingale condition. For this, we first observe that for $i \geq 1$,

$$X_i^2 = (Y_1 + \cdots + Y_i)^2 = \sum_{k=1}^{i} Y_k^2 + 2 \sum_{1\leq j<k\leq i} Y_j Y_k,$$

and therefore

$$\mathbb{E}[X_i^2 \mid Y_0, \ldots, Y_{i-1}] = \sum_{k=1}^{i} \mathbb{E}[Y_k^2 \mid Y_0, \ldots, Y_{i-1}] + 2 \sum_{1\leq j<k\leq i} \mathbb{E}[Y_j Y_k \mid Y_0, \ldots, Y_{i-1}].$$

Note that Y_k^2 is simply the constant 1, so every term in the first sum is just 1. For the second sum, note that if j and k are both less than i, then $\mathbb{E}[Y_j Y_k \mid Y_0, \ldots, Y_{i-1}] = Y_j Y_k$, by Theorem 2.9(ii). On the other hand, if $k = i$, then $j < i$ and

$$\mathbb{E}[Y_j Y_i \mid Y_0, \ldots, Y_{i-1}] = Y_j \mathbb{E}[Y_i \mid Y_0, \ldots, Y_{i-1}] = Y_j \mathbb{E}[Y_i] = 0,$$

using Theorems 2.9(iii) and 2.9(v) and the fact that $\mathbb{E}[Y_i] = 0$. Putting this together, we conclude that

$$\mathbb{E}[X_i^2 \mid Y_0, \ldots, Y_{i-1}] = i + 2 \sum_{1\leq j<k\leq i-1} Y_j Y_k.$$

However, similar reasoning shows us that

$$X_{i-1}^2 = (i-1) + 2 \sum_{1\leq j<k\leq i-1} Y_j Y_k$$

and thus that

$$\mathbb{E}[V_i \mid Y_0, \ldots, Y_{i-1}] = 2 \sum_{1\leq j<k\leq i-1} Y_j Y_k = V_{i-1},$$

which proves the martingale property.

Recall that T is the first time for which $X_i \in \{a, -b\}$. As in the previous example, let's believe that $\mathbb{E}[T] < \infty$, which is not too hard to prove. We would like to apply condition (c) to the martingale $(V_i)_{i\geq 0}$, but unfortunately, this martingale does not have bounded differences. However, there is a neat trick to get around this. Consider the stopped sequence $(V_i^T)_{i\geq 0}$. By Theorem 5.3, this process is also a martingale. Moreover, we claim that this sequence has bounded differences. Indeed, consider $|V_i^T - V_{i-1}^T|$. If $i > T$, then $V_i^T = V_{i-1}^T$, so this difference is 0. If $i \leq T$, then $|X_i| \leq \max\{a, b\}$, and so

$$|V_i^T - V_{i-1}^T| = |(X_i^2 - i) - (X_{i-1}^2 - (i-1))| \leq 1 + |X_i^2 - X_{i-1}^2| \leq 1 + 2\max\{a^2, b^2\}.$$

This is a bound independent of i, so the sequence $(V_i^T)_{i\geq 0}$ does have bounded differences, and $\mathbb{E}[T] < \infty$, so condition (c) holds. We conclude that $\mathbb{E}[V_T] = \mathbb{E}[V_0] = 0$, implying that

$$0 = \mathbb{E}[V_T] = \mathbb{E}[X_T^2 - T] = \mathbb{E}[X_T^2] - \mathbb{E}[T],$$

and thus that $\mathbb{E}[T] = \mathbb{E}[X_T^2]$. But we already know that

$$\Pr(X_T = a) = \frac{b}{a+b} \qquad \text{and} \qquad \Pr(X_T = -b) = \frac{a}{a+b}$$

implying that

$$\mathbb{E}[X_T] = \frac{b}{a+b} \cdot a^2 + \frac{a}{a+b} \cdot (-b)^2 = \frac{a^2 b + ab^2}{a+b} = ab.$$

Therefore, $\mathbb{E}[T] = ab$. Try proving this without appealing to martingale theory—it's not so easy!

4. Unbalanced coin tosses Let's go back to our frog, standing on a lily pad labeled 0, wanting to reach her home at lily pad a, and not wanting to fall into the hole at lily pad $-b$. She's *pretty* sure that her home is to the right of her current position, so she doesn't want to just jump left or right with equal probability. Instead, she jumps right with probability p and left with probability $1-q$, for some $\frac{1}{2} < p < 1$ (we assume $p > \frac{1}{2}$ because she's pretty sure her house is to the right, but the analysis works in essentially the same way for any $p \neq \frac{1}{2}$). This is equivalent to a lucky gambler, who finds a casino where the coin lands heads more frequently than tails, and thus the gambler can actually make money!

At first glance, it looks like the techniques we've developed can't say anything about what happens in this scenario. Everything we've done with martingales has been about *fair* games, and here's a patently *unfair* game. In particular, the intuition that "you can't beat the system" is obviously wrong for this game: the game itself is an instance of beating the system!

Nonetheless, the techniques we've developed are actually quite robust. Formally, let $Y_0 = 0$ and let $Y_1, Y_2, \ldots$ be independent random variables with

$$\Pr(Y_i = 1) = p, \qquad \Pr(Y_i = -1) = q = 1 - p.$$

Also, let $X_i = Y_1 + \cdots + Y_i$, representing the frog's position (or the gambler's winnings) after i rounds. Clearly, $(X_i)_{i\geq 0}$ is not a martingale, since this is not a fair game. However, let's define

$$Z_i = \left(\frac{q}{p}\right)^{X_i} = \left(\frac{q}{p}\right)^{Y_1 + \cdots + Y_i}$$

We claim that $(Z_i)_{i\geq 0}$ is a martingale with respect to $(Y_i)_{i\geq 0}$. The fact that Z_i is determined by $Y_0, \ldots, Y_i$ is clear, and the martingale property follows because

$$\begin{aligned}
\mathbb{E}[Z_i \mid Y_0, \ldots, Y_{i-1}] &= \mathbb{E}\left[\left(\frac{q}{p}\right)^{Y_1} \cdots \left(\frac{q}{p}\right)^{Y_i} \;\middle|\; Y_0, \ldots, Y_{i-1}\right] \\
&= \left(\frac{q}{p}\right)^{Y_1} \cdots \left(\frac{q}{p}\right)^{Y_{i-1}} \mathbb{E}\left[\left(\frac{q}{p}\right)^{Y_i}\right] \\
&= Z_{i-1}\left(p \cdot \left(\frac{q}{p}\right)^{1} + q \cdot \left(\frac{q}{p}\right)^{-1}\right) \\
&= Z_{i-1}(q+p) \\
&= Z_{i-1}.
\end{aligned}$$

Now, let T be the first time that $X_i \in \{a, -b\}$. We again have that $\mathbb{E}[T] < \infty$, via a skipped computation. We would like to apply condition (c) from Theorem 6.1, but for this we would need that $(Z_i)_{i\geq 0}$ has bounded differences. Unfortunately, it does not: the ratio between Z_i and Z_{i-1} is bounded, but if Z_{i-1} is already huge, then the difference between Z_i and Z_{i-1} would be huge as well. Nonetheless, we can do the same trick we did before: consider the stopped process $(Z_i^T)_{i\geq 0}$, which we know is a martingale by Theorem 5.3. When $i \geq T$ this clearly has bounded differences, as in that case $Z_i^T - Z_{i-1}^T = 0$. When $i < T$, we know that $(q/p)^a \leq Z_i^T \leq (q/p)^{-b}$, and similarly for Z_{i-1}^T, which implies that their difference is bounded by $(p/q)^b$.

Therefore, the optional stopping theorem applies, and we conclude that

$$\mathbb{E}[Z_T] = \mathbb{E}[Z_0] = 1.$$

On the other hand, we can compute that

$$\mathbb{E}[Z_T] = \Pr(X_T = a)\left(\frac{q}{p}\right)^{a} + \Pr(X_T = b)\left(\frac{q}{p}\right)^{-b}.$$

Letting $x = \Pr(X_T = a)$, so that $1 - x = \Pr(X_T = b)$, we have that

$$1 = \mathbb{E}[Z_T] = x\left(\frac{q}{p}\right)^{a} + (1-x)\left(\frac{q}{p}\right)^{-b} = \left(\frac{q}{p}\right)^{-b} + x\left(\left(\frac{q}{p}\right)^{a} - \left(\frac{q}{p}\right)^{-b}\right).$$

Solving for x shows that

$$\Pr(X_T = a) = x = \frac{(p/q)^b - 1}{(p/q)^b - (q/p)^a} = \frac{(p/q)^{a+b} - (p/q)^a}{(p/q)^{a+b} - 1}$$

Note that this formula doesn't make any sense if $p = q$ (or equivalently $p = \frac{1}{2}$), since in that case both numerator and denominator equal 0. The reason this fails is that if

$p = q$, then the random variable Z_i simply equals 1 for all i! In this case (and only this case) we do not have $\mathbb{E}[T] < \infty$, and hence we can't apply the optional stopping theorem.

We can also compute $\mathbb{E}[T]$, as follows. Let's let $A_i = X_i - i(p-q)$. Then $(A_i)_{i\geq 0}$ is again a martingale: it's clear that A_i is determined by $Y_0, \dots, Y_i$, and

$$\begin{aligned}\mathbb{E}[A_i \mid Y_0, \dots, Y_{i-1}] &= \mathbb{E}[Y_0 + \cdots + Y_i - i(p-q) \mid Y_0, \dots, Y_{i-1}] \\ &= Y_0 + \cdots + Y_{i-1} + \mathbb{E}[Y_i] - i(p-q) \\ &= X_{i-1} + (p-q) - i(p-q) \\ &= X_{i-1} - (i-1)(p-q) = A_{i-1}.\end{aligned}$$

Since this is a martingale and it has bounded increments, the optional stopping theorem applies, and we have that $\mathbb{E}[A_T] = \mathbb{E}[A_0] = 0$. Thus,

$$0 = \mathbb{E}[A_T] = \mathbb{E}[X_T] - \mathbb{E}[T] = \left(\frac{(p/q)^{a+b} - (p/q)^a}{(p/q)^{a+b} - 1} \cdot a + \frac{(p/q)^a - 1}{(p/q)^{a+b} - 1} \cdot b\right) - \mathbb{E}[X_T].$$

Rearranging, we see that

$$\mathbb{E}[X_T] = \frac{a(p/q)^{a+b} + (b-a)(p/q)^a - b}{(p/q)^{a+b} - 1}$$

5. Guessing in a deck of cards We play the following game. I shuffle a deck of cards, then start dealing them face-up on the table, one by one. Your goal is to guess one time when I will deal out a red card. Namely, at any point, you can say **STOP**, and your goal is to pick when to **STOP** to maximize the probability that the next card is red. For example, you could say **STOP** even before I begin dealing, in which case your win probability is 1/2. Alternately, you could hope that I'll deal a bunch of black cards early on, and then say **STOP**; if I really do deal a bunch of black cards that'll be helpful for you, but what if I don't?

It turns out that no matter your strategy, your odds of winning are exactly 1/2. To see this, let X_i be the proportion of the unrevealed cards (the final $52 - i$ cards) that are red. On the homework, you proved that $(X_i)_{i\geq 0}$ is a martingale. Moreover, we see that X_i is exactly the odds that the next card is red if you say **STOP** right before I deal the ith card; indeed, among the remaining $52 - i$ cards, exactly an X_i fraction of them are red, so your odds of succeeding are X_i.

Now, your stopping strategy is simply a stopping time! So we care about $\mathbb{E}[X_T]$. Moreover, T is certainly bounded (by 52), so condition (a) from Theorem 6.1 holds. So $\mathbb{E}[X_T] = \mathbb{E}[X_0]$, and we just argued that X_0 is simply 1/2. So your odds of winning are 1/2 regardless of strategy.

6. Runs of dice Suppose we keep rolling fair dice. Let $D_0 = 0$, and $D_1, D_2, D_3, \dots$ be the outcome of the die rolls, so that these are independent random variables which take

on the values $\{1, \dots, 6\}$ each with probability $1/6$. We would like to understand the average amount of time until we see three sixes in a row. Let T be the time when we first see three sixes in a row, and note that T is a stopping time with respect to the sequence $(D_i)_{i \geq 0}$.

In this example, the difficulty is cleverly coming up with a martingale that we can use. Let's define R_n to equal $1 + 6 + 6^2 + \cdots + 6^k$, if the nth die roll is the kth consecutive 6. So $R_n = 1$ if $D_n \neq 6$, whereas if $D_n = 6$ then R_n is equal to $6R_{n-1} + 1$. Finally, let's define $X_n = R_n - n$. Then we claim that $(X_i)_{i \geq 0}$ is a martingale with respect to $(D_i)_{i \geq 0}$. Indeed, it is clear that X_i is determined by $D_0, \dots, D_i$. For the martingale property, note that

$$R_n = \begin{cases} 6R_{n-1} + 1 & \text{with probability } \frac{1}{6} \\ 1 & \text{with probability } \frac{5}{6} \end{cases}$$

and thus

$$\mathbb{E}[R_n \mid D_0, \dots, D_{n-1}] = \frac{1}{6} \cdot (6R_{n-1} + 1) + \frac{5}{6} \cdot 1 = R_{n-1} + \frac{1}{6} + 56 = R_{n-1} + 1.$$

Therefore,

$$\mathbb{E}[X_n \mid D_0, \dots, D_{n-1}] = \mathbb{E}[R_n \mid D_0, \dots D_{n-1}] - n = (R_{n-1} + 1) - n = X_{n-1}.$$

Now, we again have that $\mathbb{E}[T] < \infty$, using another simple computation that I'll omit. As before, we don't actually have that $(X_i)_{i \geq 0}$ has bounded increments, but the same simple trick as before gets around this: the sequence $(X_i^T)_{i \geq 0}$ *does* have bounded increments, since $|X_i^T - X_{i-1}^T| \leq 6^3$. So we can apply the optional stopping theorem, and we conclude that $\mathbb{E}[X_T] = \mathbb{E}[X_0] = 1$, since $X_0 = R_0 - 0 = 1 - 0 = 1$. Thus,

$$1 = \mathbb{E}[X_T] = \mathbb{E}[R_T] - \mathbb{E}[T] = 1 + 6 + 6^2 + 6^3 - \mathbb{E}[T]$$

which yields

$$\mathbb{E}[T] = 6 + 6^2 + 6^3 = 258.$$

Addendum: the generalized tower property

I forgot to tell you about a very useful property of conditional expectation, which generalizes Theorem 2.9(iv). It's also called the tower property.

Theorem. *Let X, Y, Z be random variables. Then*

$$\mathbb{E}[X \mid Y] = \mathbb{E}[\mathbb{E}[X \mid Y, Z] \mid Y]. \tag{$*$}$$

Note that this recovers the earlier tower property in case Y is just a constant random variable (i.e. it doesn't matter).

Proof. First, let's think about what this equality is supposed to mean. Both sides are random variables, which depend on the value of Y. Namely, for each $y \in \mathcal{Y}$, the left-hand side of $(*)$ equals $\mathbb{E}[X \mid Y = y]$ with probability $\Pr(Y = y)$, and the right-hand side of $(*)$ equals $\mathbb{E}[\mathbb{E}[X \mid Y, Z] \mid Y = y]$ with probability $\Pr(Y = y)$. So proving that these two random variables are equal *means* proving that for each $y_0 \in \mathcal{Y}$,

$$\mathbb{E}[X \mid Y = y_0] = \mathbb{E}[\mathbb{E}[X \mid Y, Z] \mid Y = y_0]. \tag{$**$}$$

Indeed, in case we reveal the random outcome of Y and we find that $Y = y_0$, then the two sides of $(*)$ simply become the two sides of $(**)$. So the equality of random variables in $(*)$ simply *means* that $(**)$ holds for all $y_0 \in \mathcal{Y}$.

So let's prove $(**)$. We fix $y_0 \in \mathcal{Y}$. We begin with the right-hand side, and manipulate it. Recall that $\mathbb{E}[X \mid Y, Z]$ is a random variable that takes on the value $\mathbb{E}[X \mid Y = y, Z = z]$ with probability $\Pr(Y = y, Z = z)$ for all $y \in \mathcal{Y}, z \in \mathcal{Z}$. Therefore,

$$\mathbb{E}[\mathbb{E}[X \mid Y, Z] \mid Y = y_0] = \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} \mathbb{E}[X \mid Y = y, Z = z] \Pr(Y = y, Z = z \mid Y = y_0).$$

Now, what is $\Pr(Y = y, Z = z \mid Y = y_0)$? Well, if we know that $Y = y_0$, then certainly Y cannot take any value other than y_0. So in other words, we see that

$$\Pr(Y = y, Z = z \mid Y = y_0) = \begin{cases} \Pr(Z = z \mid Y = y_0) & \text{if } y = y_0 \\ 0 & \text{otherwise.} \end{cases}$$

Plugging this in, we see that

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[X \mid Y, Z] \mid Y = y_0] &= \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} \mathbb{E}[X \mid Y = y, Z = z] \Pr(Y = y, Z = z \mid Y = y_0) \\
&= \sum_{z \in \mathcal{Z}} \mathbb{E}[X \mid Y = y_0, Z = z] \Pr(Z = z \mid Y = y_0) \\
&= \sum_{z \in \mathcal{Z}} \left(\sum_{x \in \mathcal{X}} x \Pr(X = x \mid Y = y_0, Z = z) \right) \Pr(Z = z \mid Y = y_0) \\
&= \sum_{x \in \mathcal{X}} x \sum_{z \in \mathcal{Z}} \Pr(X = x \mid Y = y_0, Z = z) \Pr(Z = z \mid Y = y_0) \\
&= \sum_{x \in \mathcal{X}} x \sum_{z \in \mathcal{Z}} \Pr(X = x, Z = z \mid Y = y_0),
\end{aligned}$$

where the final step uses the definition of conditional expectation, namely that

$$\Pr(X = x \mid Y = y_0, Z = z) = \frac{\Pr(X = x, Z = z \mid Y = y_0)}{\Pr(Z = z \mid Y = y_0)}.$$

Continuing our computation, we see that

$$\begin{aligned}\mathbb{E}[\mathbb{E}[X \mid Y, Z] \mid Y = y_0] &= \sum_{x \in \mathcal{X}} x \sum_{z \in \mathcal{Z}} \Pr(X = x, Z = z \mid Y = y_0) \\ &= \sum_{x \in \mathcal{X}} x \Pr(X = x \mid Y = y_0) \\ &= \mathbb{E}[X \mid Y = y_0],\end{aligned}$$

which proves $(**)$. □

Let me remark that there is another proof besides this long and technical one, which is much more conceptual. Namely, there is an alternative definition of the conditional expectation $\mathbb{E}[X \mid Y]$, which defines it as the unique random variable satisfying certain properties. It is then quite straightforward to verify that $\mathbb{E}[\mathbb{E}[X \mid Y, Z] \mid Y]$ is a random variable satisfying those same properties, which immediately implies that $\mathbb{E}[X \mid Y] = \mathbb{E}[\mathbb{E}[X \mid Y, Z] \mid Y]$.

Unfortunately, while this step becomes very simple in this more conceptual proof, there is some amount of "preservation of difficulty". Indeed, if one takes this alternative definition of conditional expectation, then one has to *prove* that there always exists a random variable satisfying the desired properties, and that it is always unique. Those proofs essentially boil down to roughly the same kind of manipulations that we did above. Said differently, even in math, you can't beat a fair game! Sometimes proofs are just hard, and there's no way around it.