Conditional Expectations: Review and Lots of Examples

David S. Rosenberg

Abstract

The goal of this document is to get the reader to some level of proficiency in calculating and manipulating conditional expectations and variances. We're less concerned with providing definitions and theorems in their most precise and general form. The proofs for various identities are not provided for the sake of rigor, but rather because they give the opportunity to practice exactly the types of maniuplations and calculations that are the point of this document. For a small additional challenge, you can consider each theorem statement as just an exercise to complete for additional practice.

1 Basic Expectation

Let $Y \in \mathcal{Y} \subset \mathbf{R}$ be a random variable. In this document, we'll discuss taking the expectation of Y with respect to many different distributions. For simplicity, let's suppose $\mathcal{Y}$ is a finite set, although all results will hold for general random variables. Let random variable Y have a distribution described by the probability mass function (PMF) $p(y)$. Recall the definition of expectation of Y:

$$\mathbb{E}[Y] = \sum_{y \in \mathcal{Y}} yp(y).$$

Remark 1. Even though we ususally write expectations in terms of random variables, it's best to think of expectations as **properties of distributions**. Notice that the expression on the RHS[1] above makes no reference to any particular random variable. In fact, all random variables with the same PMF $p(x)$ have the same expectation. So whenever we see an expectation operator, we should be thinking about the distribution it's acting on.

[1] RHS means "right hand side", and we'll frequently use it to refer to the right hand side of an equation. LHS is defined accordingly.

2 Conditional Expectation

Let's now introduce another random element into the mix. Let $X \in \mathcal{X}$ be defined on the same probability space as Y, and let's write $p(x, y)$ for their joint PMF. Recall that the conditional distribution of Y given $X = x$ is represented by the conditional PMF

$$p(y \mid x) = \frac{p(x, y)}{p(x)}.$$

For each fixed x, $p(y \mid x)$ represents a distribution on $\mathcal{Y}$ in the sense that $p(y \mid x) \geq 0 \; \forall y \in \mathcal{Y}$ and $\sum_{y \in \mathcal{Y}} p(y \mid x) = 1$.

Definition 1. The conditional expectation of Y given $X = x$, denoted $\mathbb{E}[Y \mid X = x]$ and occasionally $\mathbb{E}[Y \mid x]$, is the expectation of the distribution represented by $p(y \mid x)$. That is,

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

As x changes, the conditional distribution of Y given $X = x$ will typically change as well, and so might the conditional expectation of Y given $X = x$. So we can view $\mathbb{E}[Y \mid X = x]$ as a function of x. To emphasize this, let's define the function $f : \mathcal{X} \to \mathbf{R}$ such that $f(x) = \mathbb{E}[Y \mid X = x]$. Note that there is nothing random about this function: the same x always gives us the same $f(x)$ as output. We can now define $\mathbb{E}[Y \mid X]$:

Definition 2. We define the **conditional expectation of** Y **given** X, denoted $\mathbb{E}[Y \mid X]$, as the **random variable** $f(X)$, where $f(x) = \mathbb{E}[Y \mid X = x]$.

In other words, $\mathbb{E}[Y \mid X]$ is what we get when we plug in the random variable X to the deterministic function $f(x)$. Since X is random, $f(X)$ and thus $\mathbb{E}[Y \mid X]$ are themselves random variables.

Remark 2. There's often a temptation to write $f(X) = \mathbb{E}[Y \mid X = X]$. Avoid this. One of the issues is that it's ambiguous: you might interpret it as conditioning on the event that $X = X$, which always occurs. It's an unfortunate notational awkwardness that one learns to work around.

Remark 3. We can generalize conditional expectation to condition on multiple random elements in the obvious way. For example, if $f(x, z) = \mathbb{E}[Y \mid X = x, Z = z]$ then $\mathbb{E}[Y \mid X, Z] = f(X, Z)$.

Exercise 1. Show that if $X \in \mathcal{X}$ has PMF $p(x)$, then $\mathbb{E}[h(X)\mathbb{E}[Y \mid X]] = \sum_{x \in \mathcal{X}} p(x)h(x)\mathbb{E}[Y \mid X = x]$.

Proof. Let $f(x) = \mathbb{E}[Y \mid X = x]$. Then

$$\begin{aligned}\mathbb{E}[h(X)\mathbb{E}[Y \mid X]] &= \mathbb{E}[h(X)f(X)] \quad \text{definition of } \mathbb{E}[Y \mid X] \\ &= \sum_{x \in \mathcal{X}} p(x)h(x)f(x) \quad \text{definition of } \mathbb{E} \\ &= \sum_{x \in \mathcal{X}} p(x)h(x)\mathbb{E}[Y \mid X = x]\,.\end{aligned}$$

□

3 Identities for conditional expectations

There are a lot of "rules" for manipulating conditional expectations, and the hope of this document is to get you comfortable with all the main ones. Here we list the rules, and in the next section we'll give some derivations and discussion. We'll give a short-hand expression for each identity, mostly borrowed from [KBH19, Ch 9] and [Wik20], so we can refer to them in derivations.

- **Adam's Law / Law of Iterated Expectation**:
  - Simple: $\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}Y$
  - More general: $\mathbb{E}[\mathbb{E}[Y \mid g(X)] \mid f(g(X))] = \mathbb{E}[Y \mid f(g(X))]$ for any f and g with compatible domains and ranges.
- **Independence:** $\mathbb{E}[Y \mid X] = \mathbb{E}[Y]$ if X and Y are independent.
- **Taking out what is known**[2]**:** $\mathbb{E}[h(X)Z \mid X] = h(X)\mathbb{E}[Z \mid X]$.
- **Linearity**: $\mathbb{E}[aX + bY \mid Z] = a\mathbb{E}[X \mid Z] + b\mathbb{E}[Y \mid Z]$, for any $a, b \in \mathbf{R}$.
- **Projection interpretation**: $\mathbb{E}[(Y - \mathbb{E}[Y \mid X])h(X)] = 0$ for any function $h : \mathcal{X} \to \mathbf{R}$.
- **Keeping just what is needed**: $\mathbb{E}[XY] = \mathbb{E}[X\mathbb{E}[Y \mid X]]$ for $X, Y \in \mathbf{R}$.

[2] This is the conditional version of $\mathbb{E}[cX] = c\mathbb{E}[X]$, for any constant $c \in \mathbf{R}$. But that is an equation of two numbers, while the conditional version is an equality of random variables. The idea is that inside the conditional expectation, we think of X as being constant, and thus $h(X)$ is also constant. As such, we can pull $h(X)$ out of the expectation. When it's on the outside of the expectation, $h(X)$ is random again.

3.1 Law of Iterated Expectations

Since $\mathbb{E}[Y \mid X]$ is a random variable, it has a distribution. What is the expectation of this distribution? In math, we can write that as $\mathbb{E}[\mathbb{E}[Y \mid X]]$. The inner expectation is over Y, and the outer expectation is over X. To clarify, this could be written as $\mathbb{E}_X[\mathbb{E}_Y[Y \mid X]]$, though this is rarely done in practice unless we need to specify the distributions that the variables are referring to, as in $\mathbb{E}_{X\sim p_1(x)}\left[\mathbb{E}_{Y\sim p_2(y|x)}[Y \mid X]\right]$.

Just like all other [unconditional] expectations, $\mathbb{E}[\mathbb{E}[Y \mid X]]$ is just a number: it's not random. It turns out, this "iterated expectation" can be written much more simply:

Theorem 1 (Law of Iterated Expectations, "Adam's Law")**.** *For random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y} \subset \mathbf{R}$ defined on the same probability space,*

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

Proof. We'll prove this for the case of finite $\mathcal{X}$ and $\mathcal{Y}$, but the result holds for arbitrary random variables. As above, let $f(x) = \mathbb{E}[Y \mid X = x]$. Then

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[Y \mid X]] &= \mathbb{E}[f(X)] = \sum_{x\in\mathcal{X}} p(x)f(x) \\
&= \sum_{x\in\mathcal{X}} p(x)\mathbb{E}[Y \mid X = x] \\
&= \sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)y \\
&= \sum_{y\in\mathcal{Y}} y\sum_{x\in\mathcal{X}} p(x,y) \\
&= \sum_{y\in\mathcal{Y}} yp(y) \\
&= \mathbb{E}Y
\end{aligned}$$

□

Exercise 2. Let $\mathbb{1}[W = 1]$ denote the random variable that takes the value 1 if $W = 1$ and 0 otherwise. Show that $\mathbb{E}[\mathbb{1}[W = 1]Y] = \mathbb{P}(W = 1)\mathbb{E}[Y \mid W = 1]$.

Proof. We have

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}\left[W=1\right]Y\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left[W=1\right]Y \mid W\right]\right] && \text{by Law if Iterated Expectations} \\
&= \mathbb{E}\left[\mathbb{1}\left[W=1\right]\mathbb{E}\left[Y \mid W\right]\right] && \text{taking out what is known} \\
&= \mathbb{P}\left(W=1\right)\mathbb{1}\left[1=1\right]\mathbb{E}\left[Y \mid W=1\right] && \text{Exercise 1} \\
&\quad +\mathbb{P}\left(W=0\right)\underbrace{\mathbb{1}\left[0=1\right]}_{\equiv 0}\mathbb{E}\left[Y \mid W=0\right] \\
&= \mathbb{P}\left(W=1\right)\mathbb{E}\left[Y \mid W=1\right]
\end{aligned}$$

□

3.1.1 Information processing

We'll show later that $\mathbb{E}\left[Y \mid X\right]$ is the best prediction we can make for Y given X (in terms of mean squared error). What if we have some function $f : \mathcal{X} \to \mathcal{X}'$ and we consider $\mathbb{E}\left[Y \mid f(X)\right]$. Does $f(X)$ have more, less, or the same information as X? Well, it could have much less, such as if $f(x) \equiv 0$ for any x. If $f(x)$ is injective (i.e. if $x \neq y$ then $f(x) \neq f(y)$), then $f(X)$ has the same information as X, since we can always recover X from $f(X)$ by $X = f^{-1}\left(f(X)\right)$. So in some sense, $f(X)$ has at most as much information[3] as X. So generally speaking, $\mathbb{E}\left[Y \mid f(X)\right]$ will not be as good a prediction of Y as $\mathbb{E}\left[Y \mid X\right]$.

We'll now discuss the more general form of Adam's Law presented above. Suppose we have an information processing chain: $x \mapsto g(x) \mapsto f(g(x))$. We can think $g(X)$ as a "processed" or "coarsened" version of X. So $\mathbb{E}\left[Y \mid g(X)\right]$ is our best approximation for Y given $g(X)$. Suppose we have $f(g(X))$, which is an even more processed version of X, and we want the best prediction for $\mathbb{E}\left[Y \mid g(X)\right]$ given only $f(g(X))$. It turns out, that prediction is also the best prediction for Y given only $f(g(X))$. In math:

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

No proof right now, but we'll have another interpretation in terms of projections in the next section.

Remark 4. If we take $g(x,z) = (x,z)$ and $f(g(x,z)) = z$ in the generalized Adam's Law, we that

$$\mathbb{E}\left[\mathbb{E}\left[Y \mid X,Z\right] \mid Z\right] = \mathbb{E}\left[Y \mid Z\right],$$

which is often useful in practice.

[3] These notions are formalized in information theory by the **data processing inequality** (see e.g. [CT06, Chapter 2]), but we're just looking for intuition here, so we don't need to be formal.

3.2 Projection interpretation

As exercises in using our other identities, in this section we'll prove the "projection interpretation" and that $\mathbb{E}[Y \mid X]$ gives the best possible prediction for Y based only on X. We'll also discuss why this allows us to characterize $\mathbb{E}[Y \mid X]$ as a projection of the random variable Y onto the space of random variables that depend only on X.

3.2.1 What we can say about residuals

If we think of $\mathbb{E}[Y \mid X]$ as a prediction for Y given X, then $Y - \mathbb{E}[Y \mid X]$ is the **residual** of that prediction. The next theorem shows that the residual for $\mathbb{E}[Y \mid X]$ is "orthogonal" to every random variable of the form $h(X)$. In the corollary that follows, we'll relate orthogonality to covariance and correlation.

Theorem 2 (Projection interpretation)**.** *For any $h : \mathcal{X} \to \mathbf{R}$,* $\mathbb{E}[(Y - \mathbb{E}[Y \mid X])h(X)] = 0$.

Proof. We have

$$\begin{aligned}
&\mathbb{E}[(Y - \mathbb{E}[Y \mid X])h(X)] \\
&= \mathbb{E}[Yh(X)] - \mathbb{E}[\mathbb{E}[Y \mid X]\, h(X)] \qquad \text{by linearity} \\
&= \mathbb{E}[Yh(X)] - \mathbb{E}[\mathbb{E}[Yh(X) \mid X]] \qquad \text{taking out what is known (in reverse)} \\
&= \mathbb{E}[Yh(X)] - \mathbb{E}[Yh(X)] \qquad \text{Adam's Law} \\
&= 0
\end{aligned}$$

□

Definition 3. The **covariance** of random variables X and Y is defined by $\text{Cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y) = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y$. If $\text{Cov}(X, Y) = 0$, then we say X and Y are **uncorrelated**.

Corollary 1. *The residual $Y - \mathbb{E}[Y \mid X]$ and $h(X)$ are uncorrelated (i.e. have covariance 0) for every function $h : \mathcal{X} \to \mathbf{R}$.*

Proof. Note that $\mathbb{E}[Y - \mathbb{E}[Y \mid X]] = \mathbb{E}Y - \mathbb{E}[\mathbb{E}[Y \mid X]] = 0$ by linearity and Adam's Law. So

$$\begin{aligned}
\text{Cov}(Y - \mathbb{E}[Y \mid X], h(X)) &= \mathbb{E}[(Y - \mathbb{E}[Y \mid X])\, h(X)] - \underbrace{\mathbb{E}[Y - \mathbb{E}[Y \mid X]]}_{=0}\mathbb{E}h(X) \\
&= \mathbb{E}[(Y - \mathbb{E}[Y \mid X])\, h(X)] = 0
\end{aligned}$$

where the last equality is by Theorem 2. □

Remark 5. Note that Corollary 1 speaks about correlation, but not independence! For example, the residual $Y - \mathbb{E}[Y \mid X]$ may have more variance for some values of X than others. Thus $Y - \mathbb{E}[Y \mid X]$ is generally not independent of X, even though it is uncorrelated with every random variable of the form $h(X)$. [Why are we saying $h(X)$ here instead of just X? X is not necessarily real-valued, and covariance and correlation are defined specifically for random variables. Independence is defined for any type of random element. If X is a random variable (i.e. real-valued), then we can certainly say that $Y - \mathbb{E}[Y \mid X]$ and X are uncorrelated.]

3.2.2 Conditional expectation gives the best prediction

We now use Theorem 2 to prove that conditional expectation gives the best possible prediction of Y based on X.

Theorem 3 (Conditional expectation minimizes MSE)**.** *Suppose $X \in \mathcal{X}$ and $Y \in \mathbf{R}$ on the same probability space. Let $g(x) = \mathbb{E}[Y \mid X = x]$. Then*

$$g(x) = \arg\min_f \mathbb{E}\left(Y - f(X)\right)^2.$$

Proof. We have

$$\begin{aligned}
&\mathbb{E}\big[(f(X) - Y)^2\big] \\
&\quad = \mathbb{E}\big[f(X) - \mathbb{E}[Y \mid X] + \mathbb{E}[Y \mid X] - Y\big]^2 \\
&\quad = \mathbb{E}(f(X) - \mathbb{E}[Y|X])^2 + \mathbb{E}\left[(\mathbb{E}[Y|X] - Y)^2\right] \\
&\qquad + \underbrace{\mathbb{E}\left[\left(\underbrace{f(X) - \mathbb{E}[Y \mid X]}_{\text{function of } X}\right)\left(\underbrace{\mathbb{E}[Y \mid X] - Y}_{\text{residual}}\right)\right]}_{=0} \qquad \text{Projection interpretation} \\
&\quad = \mathbb{E}(f(X) - \mathbb{E}[Y|X])^2 + \mathbb{E}\left[(\mathbb{E}[Y|X] - Y)^2\right].
\end{aligned}$$

The second term in the last expression is independent of f, and the first term in the last expression is clearly minimized by taking $f(x) = \mathbb{E}[Y \mid X = x]$. □

As we'll explain below, this theorem is what justifies calling $\mathbb{E}[Y \mid X]$ a projection.

3.2.3 A variance decomposition

Sometimes it's helpful to think of decomposing Y as

$$Y = \underbrace{\mathbb{E}[Y \mid X]}_{\text{best prediction for } Y \text{ given } X} + \underbrace{Y - \mathbb{E}[Y \mid X]}_{\text{residual}}.$$

Note that the two terms on the RHS are uncorrelated, by the projection interpretation (Corollary 1). Since variance is additive for uncorrelated random variables (i.e. if X and Y are uncorrelated, then $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$), we get the following theorem:

Theorem 4 (Variance decomposition with projection)**.** *For any random $X \in \mathcal{X}$ and random variable $Y \in \mathbf{R}$, we have*

$$\text{Var}(Y) = \text{Var}(Y - \mathbb{E}[Y \mid X]) + \text{Var}(\mathbb{E}[Y \mid X]).$$

Remark 6. Theorem 3 tells us that $\mathbb{E}[Y \mid X]$ is the best approximation of Y we can get from X. We can also think of $\mathbb{E}[Y \mid X]$ as a "less random" version of Y, since $\text{Var}(\mathbb{E}[Y \mid X]) \leq \text{Var}(Y)$ [this follows immediately from the previous Theorem since variance is always ≥ 0]. We can say that $\mathbb{E}[Y \mid X]$ only keeps the randomness in Y that is predictable from X.... Why do we say this? $\mathbb{E}[Y \mid X]$ is a deterministic function of X, so there's no other source of randomness in $\mathbb{E}[Y \mid X]$.

3.2.4 [Optional] Why do we call this the "projection interpretation"?

One can consider the space of all random variables with finite variance as an inner product space with inner product given by

$$\langle X, Y \rangle = \mathbb{E}[XY]$$

and norm given by $\|Y\|^2 = \langle Y, Y \rangle = \mathbb{E}Y^2$. A random variable S' is called a **projection** (or L_2-projection) of Y onto $\mathcal{S}$ if $S' \in \mathcal{S}$ and

$$\mathbb{E}(Y - S')^2 \leq \mathbb{E}(Y - S)^2 \quad \forall S \in \mathcal{S}.$$

In words, S' is the projection of Y onto $\mathcal{S}$ if it is the best approximation of Y in $\mathcal{S}$ in terms of mean squared error (MSE). In Theorem 3 above we exactly proved that $\mathbb{E}[Y \mid X]$ is the function of X that has the smallest possible MSE for predicting Y. Thus $\mathbb{E}[Y \mid X]$ is the projection of Y onto the set of random variables $\{h(X) \mid h \text{ is any real-valued function}\}$.

Remark 7. The projection interpretation gives another way to think about the generalized Adam's Law: $\mathbb{E}\left[\mathbb{E}\left[Y \mid g(X)\right] \mid f(g(X))\right] = \mathbb{E}\left[Y \mid f(g(X))\right]$ for any f and g with compatible domains and ranges. We can think of the LHS as a sequence of two projections, while the RHS is a single projection. Adam's Law says they're equivalent. In more detail, $\mathbb{E}\left[Y \mid g(X)\right]$ is the projection of Y onto $\{h(g(X)) \mid \forall h\}$, the set of all functions of $g(X)$, and $\mathbb{E}\left[\mathbb{E}\left[Y \mid g(X)\right] \mid f(g(X))\right]$ is the projection of $\mathbb{E}\left[Y \mid g(X)\right]$ onto $\{h(f(g(X))) \mid \forall h\}$, the set of all functions of $f(g(X))$. Note that the second set of functions is a subset of the first, i.e. $\{h(f(g(X))) \mid \forall h\} \subseteq \{h(g(X)) \mid \forall h\}$, since $f(\cdot)$ may discard information from $g(X)$. So Adam's Law is saying that if we project onto a set and then project onto a subset of the original set, then we get the same thing as if we had projected Y directly onto the subset to begin with. Perhaps you can visualize this by picturing projecting a vector in $\mathbf{R}^3$ onto a 2-dimensional subspace, and then projecting the projection onto a 1-dimensional subspace contained in the 2-dimensional subspace.

3.2.5 Empirical example of the variance decomposition

To illustrate some of the concepts of the variance decomposition, let's consider the following joint distribution of (X, Y):

$$\begin{aligned} X &\sim \text{Unif}[0, 6] \\ Y \mid X = x &\sim \mathcal{N}\left(6 + 1.3\sin(x), \left[.3 + \frac{1}{4}\left|3 - x\right|\right]^2\right) \end{aligned}$$

So given $X = x$, the best predictor for Y in MSE is $\mathbb{E}\left[Y \mid X = x\right] = 6 + 1.3\sin(x)$. Figure 1 shows a sample of size $n = 300$ from this distribution. For each sampled point (x, y), we also plot $(x, \mathbb{E}\left[Y \mid X = x\right])$, which is the best prediction of Y given just $X = x$, along with the residual of that prediction. Note that the residuals hover around 0. Indeed, we should expect that since

$$\begin{aligned} &\mathbb{E}\left[Y - \mathbb{E}\left[Y \mid X\right] \mid X = x\right] \\ &= \mathbb{E}\left[Y \mid X = x\right] - \mathbb{E}\left[\mathbb{E}\left[Y \mid X\right] \mid X = x\right] \quad \text{by linearity} \\ &= \mathbb{E}\left[Y \mid X = x\right] - \mathbb{E}\left[Y \mid X = x\right]\mathbb{E}\left[1 \mid X = x\right] \quad \text{taking out what is known} \\ &= 0. \end{aligned}$$

By the variance decomposition in terms of projection (Theorem 4), we know $\text{Var}\left(Y\right) = \text{Var}\left(Y - \mathbb{E}\left[Y \mid X\right]\right) + \text{Var}\left(\mathbb{E}\left[Y \mid X\right]\right)$. Using standard variance estimators with our observed sample, we find $\widehat{\text{Var}}(Y - \mathbb{E}\left[Y \mid X\right]) \approx 0.53$, $\widehat{\text{Var}}(\mathbb{E}\left[Y \mid X\right]) \approx$

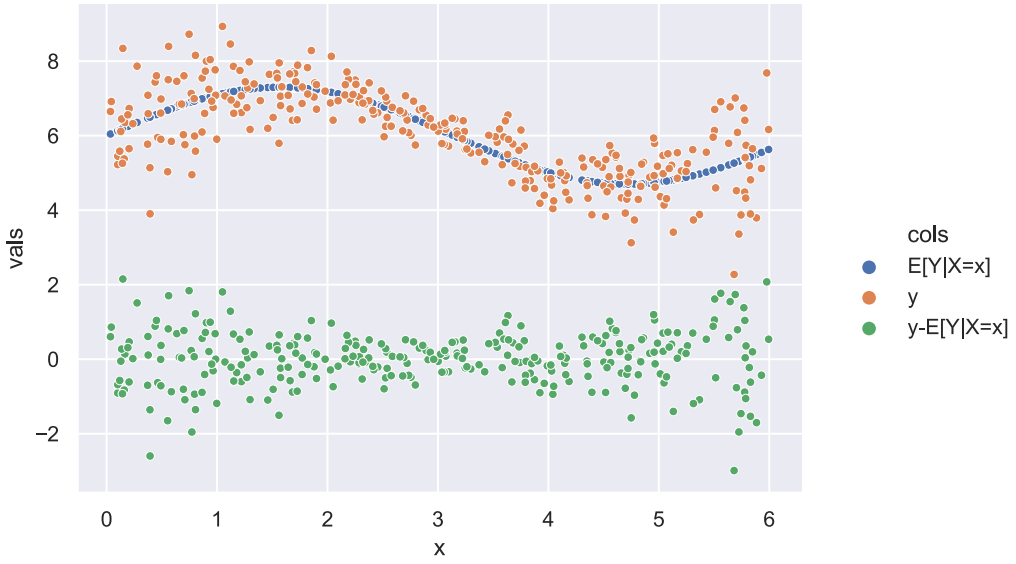

Fig. 1: This plot shows the sampled (x, y) pairs, along with the conditional expectation and residual for each: $(x, \mathbb{E}[Y \mid X = x])$ and $(x, y - \mathbb{E}[Y \mid X = x])$.

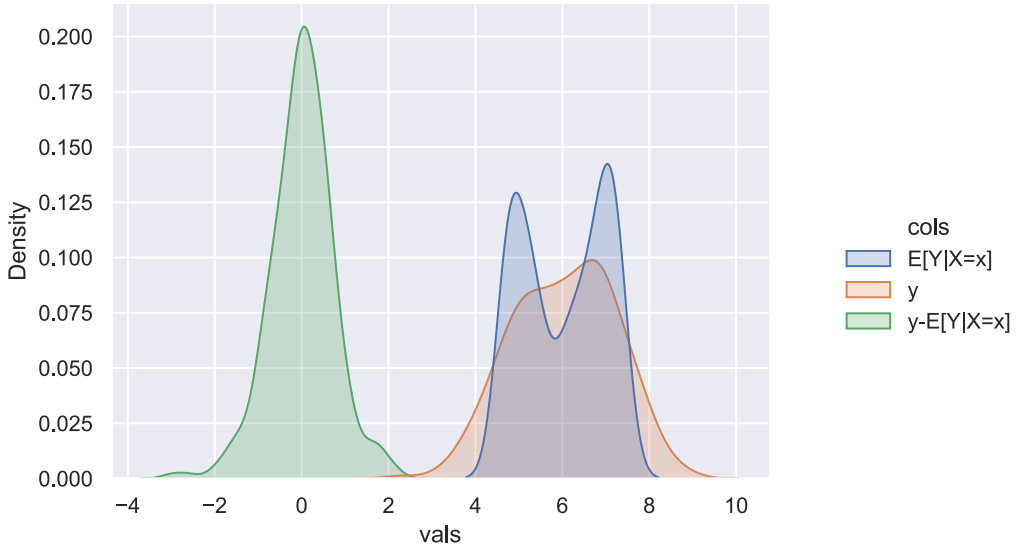

Fig. 2: This plot shows estimates of the densities of Y, $\mathbb{E}[Y \mid X]$, and $Y - \mathbb{E}[Y \mid X]$.

0.91, and $\widehat{\text{Var}}(Y) \approx 1.39$. While $\widehat{\text{Var}}(Y - \mathbb{E}[Y \mid X]) + \widehat{\text{Var}}(\mathbb{E}[Y \mid X]) = 1.43$. The gap between 1.43 and 1.39 is attributable to sampling error and vanishes as we take the sample size $n \to \infty$. In Figure 2 we show kernel density estimates of each of the distributions in the variance decomposition.

3.3 Keeping just what is needed

Theorem 5 (Keeping just what is needed)**.** *For any random variables* $X, Y \in \mathbf{R}$*,* $\mathbb{E}[XY] = \mathbb{E}[X\mathbb{E}[Y \mid X]]$.

One way to think about this is that for the purposes of computing $\mathbb{E}[XY]$, we only care about the randomness in Y that is predictable from X.

Proof. We can show this using the projection interpretation: □

$$\begin{aligned}
\mathbb{E}[XY] &= \mathbb{E}\left[X\left(\mathbb{E}[Y \mid X] + \underbrace{Y - \mathbb{E}[Y \mid X]}_{\text{residual uncorrelated with } X}\right)\right] \\
&= \mathbb{E}[X\mathbb{E}[Y \mid X]] \\
&\quad + \cancelto{0}{\mathbb{E}[X(Y - \mathbb{E}[Y \mid X])]} \quad \text{Projection interpretation} \\
&= \mathbb{E}[X\mathbb{E}[Y \mid X]]
\end{aligned}$$

Exercise 3. Give an alternative proof of $\mathbb{E}[XY] = \mathbb{E}[X\mathbb{E}[Y \mid X]]$ using Adam's Law and Taking out what is known.

Let's put Theorem 5 in a slightly more general context and consider $\mathbb{E}[g(X)h(Y)]$. Theorem 5 tells us that we get the same result if we replace $h(Y)$ by an approximation to $h(Y)$, namely $\mathbb{E}[h(Y) \mid g(X)]$. By Theorem 3, this is actually the best approximation for $h(Y)$ given $g(X)$. Can we also get the same answer if we replace $h(Y)$ by another approximation $\mathbb{E}[h(Y) \mid X]$? This approximation is potentially better than $\mathbb{E}[h(Y) \mid g(X)]$, since there may be more information in X than in $g(X)$. In the following Exercise, show that we get the same result even if we plug in the better approximation:

Exercise 4. $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)\mathbb{E}[h(Y) \mid X]]$. (Hint: You can either use the projection interpretation approach we used for the proof of Theorem 5, or it's basically a two-liner with the application of Adam's Law and Taking out what is known.

Exercise 5. Show that $\mathbb{E}[X\mathbb{E}[Y \mid Z]] = \mathbb{E}[\mathbb{E}(X \mid Z)\mathbb{E}[Y \mid Z]] = \mathbb{E}[\mathbb{E}[X \mid Z]Z]$. (This property is sometimes referred to as "self-adjointness".)

Proof. We have

$$\begin{aligned}
\mathbb{E}[X\mathbb{E}[Y \mid Z]] &= \mathbb{E}[\mathbb{E}(X\mathbb{E}[Y \mid Z] \mid Z)] \quad \text{Adam's Law} \\
&= \mathbb{E}[\mathbb{E}(X \mid Z)\mathbb{E}[Y \mid Z]] \quad \text{Taking out what is known.}
\end{aligned}$$

□

Exercise 6. Give a new proof of the "projection interpretation" (Theorem 2) using "keeping just what is needed" (Theorem 5).

3.4 Intuition Builders and Extra Exercises

Suppose $\mathbb{E}[Y \mid X] = c$ is a constant. This means that whatever information we learn from X, our best prediction for Y never changes. Does this mean that X and Y are independent? No way! For example, the variance of Y can change dramatically as a function of X, even if the expected value of Y is constant. However, if X is a real-valued random variable, we can say something about the **correlation** of X and Y.

Exercise 7. [KBH19, Ch. 9 Exercise 29] If X and Y are random variables and $\mathbb{E}[Y \mid X] = c$, then show that X and Y are uncorrelated. (Hint: It's sufficient to show that $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$.)

Proof. We have:

$$\begin{aligned}
\mathbb{E}[XY] &= \mathbb{E}[\mathbb{E}[YX \mid X]] \\
&= \mathbb{E}[X\mathbb{E}[Y \mid X]] \\
&= c\mathbb{E}[X] \\
\mathbb{E}[Y] &= \mathbb{E}[\mathbb{E}[Y \mid X]] = c \\
\text{Cov}(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = c\mathbb{E}[X] - c\mathbb{E}[X]] = 0.
\end{aligned}$$

□

Exercise 8. [KBH19, Ch. 9 Exercise 30]If X and Y are independent random variables, then we know that $\mathbb{E}[Y \mid X] = \mathbb{E}[Y]$, which is a constant. However, if we only know that X and Y are uncorrelated, then $\mathbb{E}[Y \mid X]$ is not necessarily a constant. Give an example of this. (Hint: Your job here is to come up with a joint distribution of X and Y and show it has the required properties. There are many ways to do this, so try to keep things simple. For example, you can define Y to be a deterministic function of X and keep $\mathcal{X}$ to be a small set.

Solution 1. Take $(X, Y) \in \{(-1, 1), (0, 0), (1, 1)\}$ with equal probability. Then the covariance of X and Y is 0 and $\mathbb{E}[Y \mid X = x] = \mathbb{1}[x \in \{-1, 1\}]$.

Exercise 9. We know that if X and Y are independent random variables, then $\mathbb{E}[Y \mid X] = \mathbb{E}[Y]$. But if there's another random variable W in the picture, can we also say that $\mathbb{E}[Y \mid X, W] = \mathbb{E}[Y \mid W]$? Is there a rule that we might call "Drop what is independent from the conditioning"?

Solution 2. Nope! Consider X, W i.i.d. with uniform distributions on $\{0, 1\}$. Suppose $Y = X \oplus W$. That is $Y = \mathbb{1}[X \neq W]$. Then X alone gives no information about Y. Similarly W alone gives no information about Y. Thus Y is independent of X and Y is independent of W. But Y is not independent of (X, W). In any case $\mathbb{E}[Y \mid W] = \mathbb{E}[Y] = 0.5$, while $\mathbb{E}[Y \mid X, W] = \mathbb{1}[X \neq W]$.

Exercise 10. [KBH19, Ch 9 Exercise 40]Let X_1, X_2, Y be random variables and let $A = \mathbb{E}[Y \mid X_1]$ and $B = \mathbb{E}[Y \mid X_1, X_2]$. Show that $\text{Var}(A) \leq \text{Var}(B)$.

At first glance, this result may seem counter to intuition. Usually we think that getting more information (e.g. X_1 and X_2 rather than just X_1) should reduce uncertainty, rather than increase it. Why would variance be increasing when we add more information? The devil's in the details. Here we're not talking about the uncertainty in our estimate for Y (that would be something like $\text{Var}(Y \mid X_1, X_2)$), but rather how much our estimates for Y change as we get different random X's. The more information we can use to estimate Y, the more potential there is for variation in those estimates.

Proof. We first note that $A = \mathbb{E}[B \mid X_1]$, by the generalized Adam's Law. By the projection interpretation, $B - \mathbb{E}[B \mid X_1]$ and $A = \mathbb{E}[B \mid X_1]$ are uncorrelated. Thus from

$$B = B - \mathbb{E}[B \mid X_1] + A$$

we get

$$\text{Var}(B) = \text{Var}(B - \mathbb{E}[B \mid X_1]) + \text{Var}(A).$$

Since we always have variance ≥ 0, we must have $\text{Var}(B) \geq \text{Var}(A)$.

□

Exercise 11. [KBH19, Ch 9, Exercise 41]Show that for any X and Y,

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

Proof. Let $f(x) = \mathbb{E}[Y \mid X = x]$. So $f(X)$ is our best approximation to Y given X. So

$$\begin{aligned}
\mathbb{E}[Y \mid \mathbb{E}[Y \mid X]] &= \mathbb{E}[Y \mid f(X)] \\
&= \mathbb{E}[\mathbb{E}[Y \mid X] \mid f(X)] \quad \text{generalized Adam's} \\
&= \mathbb{E}[f(X) \mid f(X)] \\
&= f(X) \quad \text{Taking out what is known}
\end{aligned}$$

□

4 Conditional variance

We could define $\text{Var}(Y|X)$ using the same approach that we used to define $\mathbb{E}[Y \mid X]$. Let $g(x) = \text{Var}(Y \mid X = x)$, where $\text{Var}(Y \mid X = x)$ is the variance of the conditional distribution $Y \mid X = x$, which is just a number. And then define $\text{Var}(Y \mid X) = g(X)$. We can also just define conditional variance directly in terms of conditional expectations:

Definition 4. The **conditional variance** of Y given X is

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

4.1 Law of Total Variance

According to wikipedia, the law of total variance goes by many names, including the variance decomposition formula, conditional variance formula, law of iterated variances, and Eve's law.

Theorem (Eve's Law)**.** *If X and Y are random variables on the same probability space, then*

$$\textit{Var}(Y) = \mathbb{E}[\textit{Var}(Y \mid X)] + \textit{Var}(\mathbb{E}[Y \mid X]).$$

On the RHS, if we write E for expectation and V for variance, the sequence of operations is EVVE. That's why this is sometimes called "Eve's law". Not a bad way to remember this important decomposition.

Let's interpret this theorem in the case that X takes values in a finite set $\mathcal{X} = \{x_1, \ldots, x_N\}$. We can call $\text{Var}(Y \mid X = x)$ the **within group** variance for the group $X = x$, and so $\mathbb{E}[\text{Var}(Y \mid X)]$ is the [weighted] average of the within group variances. This is clear just from writing out the expectation:

$$\mathbb{E}[\text{Var}(Y \mid X)] = \sum_{x \in \mathcal{X}} p(x)\text{Var}(Y \mid X = x) .$$

We can call $\text{Var}(\mathbb{E}[Y \mid X])$ the **between group** variance, where each group x is represented by the single number $\mathbb{E}[Y \mid X = x]$. If the groups have equal probabilities $p(x_1) = \cdots = p(x_N)$, then $\text{Var}(\mathbb{E}[Y \mid X])$ is just the variance of the numbers $\mathbb{E}[Y \mid X = x_1], \ldots, \mathbb{E}[Y \mid X = x_N]$. More generally, $\text{Var}(\mathbb{E}[Y \mid X])$ is the variance of the distribution described by the following table:

Probability	Value
$p(x_1)$	$\mathbb{E}[Y \mid X = x_1]$
$\vdots$	$\vdots$
$p(x_N)$	$\mathbb{E}[Y \mid X = x_N]$

Proof. Expanding the definitions:

$$\begin{aligned}
\mathbb{E}[\text{Var}(Y \mid X)] &= \mathbb{E}\left[\mathbb{E}\left[Y^2 \mid X\right] - (\mathbb{E}[Y \mid X])^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y^2 \mid X\right]\right] - \mathbb{E}\left[(\mathbb{E}[Y \mid X])^2\right] \quad \text{by linearity} \\
&= \mathbb{E}Y^2 - \mathbb{E}\left[(\mathbb{E}[Y \mid X])^2\right] \quad \text{by Adam's Law} \qquad (1)
\end{aligned}$$

and

$$\begin{aligned}
\text{Var}(\mathbb{E}[Y \mid X]) &= \mathbb{E}(\mathbb{E}[Y \mid X])^2 - (\mathbb{E}[\mathbb{E}[Y \mid X]])^2 \quad \text{def of Var} \\
&= \mathbb{E}(\mathbb{E}[Y \mid X])^2 - (\mathbb{E}Y)^2 \quad \text{by Adam's Law.}
\end{aligned}$$

Adding these expression together, we get the result. □

Remark 8. It's tempting to say that getting new information about Y from observing $X = x$ would decrease the variance. That is, it seems reasonable that $\text{Var}(Y \mid X = x) \leq \text{Var}(Y)$ for all x. But this is not the case. For example, we could have $\text{Var}(Y \mid X = x)$ very large for a particular x, but if $X = x$ is very rare, the overall variance of Y could still be much smaller. On the other hand, it is true that $\text{Var}(Y \mid X = x) \leq \text{Var}(Y)$ **on average** over X. More precisely:

$$\mathbb{E}[\text{Var}(Y \mid X)] \leq \text{Var}(Y).$$

This follows immediately from Eve's Law (Theorem 4.1) and the fact that $\text{Var}(\mathbb{E}[Y \mid X]) \geq 0$.

We can equate Eve's Law with our variance decomposition in terms of projection (Theorem 4) to get the following theorem:

Theorem 6. *If X and Y are random variables on the same probability space, then*

$$\mathbb{E}[\mathit{Var}(Y \mid X)] = \mathit{Var}(Y - \mathbb{E}[Y \mid X]) = \mathbb{E}(Y - \mathbb{E}[Y \mid X])^2$$

Proof. As an exercise in conditional expectations, we'll prove this without using Eve's Law:

Since $Y - \mathbb{E}[Y \mid X]$ has mean 0,

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

Since $\mathbb{E}[Y \mid X]$ is a function of X, we can use the generalized form of "keeping just what is needed" (Exercise 4). We have $\mathbb{E}[Y\mathbb{E}[Y \mid X]] = \mathbb{E}[\mathbb{E}[Y \mid X]\mathbb{E}[Y \mid X]]$, where we've replaced Y in the first expectation by $\mathbb{E}[Y \mid X]$. Putting it together, we get

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

where the last equality is from Equation 1 in the proof of Eve's Law above. □

Exercise 12. Suppose $A \in \mathcal{A}$ has probability mass function $\pi(a)$, for $a \in \mathcal{A} = \{1, \ldots, k\}$ and $R \in \mathcal{R}$ is an **independent** random element. Show that

$$\mathbb{E}[f(R, A)g(A)] = \frac{1}{k}\sum_{a=1}^{k} \mathbb{E}[f(R, a)]\,\pi(a)g(a).$$

Proof. In the context that this exercise arises, we start with the RHS and we need to "discover" the LHS. So starting with the LHS would be a "guess and check" approach. We'll start with the RHS:

$$\begin{aligned}
&\frac{1}{k}\sum_{a=1}^{k} \mathbb{E}[f(R, a)]\,\pi(a)g(a) \\
&= \frac{1}{k}\sum_{a=1}^{k} \pi(a)\mathbb{E}[f(R, a)g(a)] \quad \text{since } g(a) \text{ is constant} \\
&= \frac{1}{k}\sum_{a=1}^{k} \pi(a)\underbrace{\mathbb{E}[f(R, A)g(A) \mid A = a]}_{=h(a)} \quad \text{since } R \text{ and } A \text{ are independent} \\
&= \mathbb{E}[h(A)] \\
&= \mathbb{E}[\mathbb{E}[f(R, A)g(A) \mid A]] \\
&= \mathbb{E}[f(R, A)g(A)]
\end{aligned}$$

As we get more comfortable with conditional expectations, we can skip the step involving $h(a)$. □

5 Law of total covariance / Conditional covariance

First, recall the definition of the covariance of X and Y: $\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)$.

Exercise 13. Show that covariance is not affected by changing the means of the random variables. To be precise, if $X' = X + c_1$ and $Y' = Y + c_2$ for constants $c_1, c_2 \in \mathbf{R}$, then $\text{Cov}(X', Y') = \text{Cov}(X, Y)$.

Definition 5. The **conditional covariance** of X and Y given Z is

$$\begin{aligned} \text{Cov}(X, Y \mid Z) &= \mathbb{E}[(X - \mathbb{E}[X \mid Z])(Y - \mathbb{E}[Y \mid Z]) \mid Z] \\ &= \mathbb{E}[XY \mid Z] - \mathbb{E}[X \mid Z]\mathbb{E}[Y \mid Z] \end{aligned}$$

Exercise 14. Use the rules we developed above to show that the two expressions for $\text{Cov}(X, Y \mid Z)$ are equivalent.

$$\begin{aligned} \text{Cov}(X, Y \mid Z) &= \mathbb{E}[(X - \mathbb{E}[X \mid Z])(Y - \mathbb{E}[Y \mid Z]) \mid Z] && \text{definition} \\ &= \mathbb{E}[XY \mid] Z + \mathbb{E}[\mathbb{E}[X \mid Z]\mathbb{E}[Y \mid Z] \mid Z] && \text{linearity} \\ &\quad -\mathbb{E}[X\mathbb{E}[Y \mid Z] \mid Z] - \mathbb{E}[Y\mathbb{E}[X \mid Z] \mid Z] \\ &= \mathbb{E}[XY \mid Z] + \mathbb{E}[X \mid Z]\mathbb{E}[Y \mid Z]\cancelto{1}{\mathbb{E}[1 \mid Z]} && \text{taking out what is known} \\ &\quad -\mathbb{E}[Y \mid Z]\mathbb{E}[X \mid Z] - \mathbb{E}[X \mid Z]\mathbb{E}[Y \mid Z] \\ &= \mathbb{E}[XY \mid Z] - \mathbb{E}[X \mid Z]\mathbb{E}[Y \mid Z] \end{aligned}$$

Theorem 7 (Law of Total Covariance (ECCE))**.** *Suppose X and Y are random variables and Z is a random element on the same probability space. Then*

$$\textit{Cov}(X, Y) = \mathbb{E}[\textit{Cov}(X, Y \mid Z)] + \textit{Cov}(\mathbb{E}[X \mid Z], \mathbb{E}[Y \mid Z]).$$

Note: Following [KBH19, Ch 9, Exercise 43], we'll use ECCE as a shorthand for the law of total covariance, based on the sequence of expectations and covariances in the formula. (Again, also a good mnemonic.)

Proof. We have

$$\begin{aligned} \mathbb{E}[\text{Cov}(X, Y \mid Z)] &= \mathbb{E}[\mathbb{E}[XY \mid Z] - \mathbb{E}[X \mid Z]\mathbb{E}[Y \mid Z]] && \text{def} \\ &= \mathbb{E}[XY] - \mathbb{E}[\mathbb{E}[X \mid Z]\mathbb{E}[Y \mid Z]] && \text{linearity and Adam's} \end{aligned}$$

and

$$\begin{aligned}\operatorname{Cov}\left(\mathbb{E}\left[X \mid Z\right], \mathbb{E}\left[Y \mid Z\right]\right) &= \mathbb{E}\left[\mathbb{E}\left[X \mid Z\right]\mathbb{E}\left[Y \mid Z\right]\right] \\ &\quad -\mathbb{E}\left[\mathbb{E}\left[X \mid Z\right]\right]\mathbb{E}\left[\mathbb{E}\left[Y \mid Z\right]\right] \quad \text{def} \\ &= \mathbb{E}\left[\mathbb{E}\left[X \mid Z\right]\mathbb{E}\left[Y \mid Z\right]\right] - \mathbb{E}X\mathbb{E}Y \quad \text{Adam's}\end{aligned}$$

Adding these expressions together, we get

$$\mathbb{E}\left[XY\right] - \mathbb{E}X\mathbb{E}Y = \operatorname{Cov}\left(X, Y\right).$$

□

References

[CT06] Thomas M. Cover and Joy A. Thomas, *Elements of information theory (wiley series in telecommunications and signal processing)*, Wiley-Interscience, USA, 2006.

[KBH19] Joseph K. Blitzstein and Jessica Hwang, *Introduction to probability second edition*, 2nd ed., Chapman and Hall/CRC, 2019.

[Wik20] Wikipedia contributors, *Conditional expectation — Wikipedia, the free encyclopedia*, 2020, [Online; accessed 31-December-2020].