Conditional Expectation

Janko Gravner

MAT/STA 235B

January 19, 2024

Conditional expectation

Definition

Let $(\Omega, \mathcal{F}, P)$ be a probability space and X a random variable on this space such that $E|X| < \infty$. Let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra. We define a *conditional expectation* $E[X \mid \mathcal{G}]$ of X given $\mathcal{G}$ to a be a random variable Y such that:

1. Y is $\mathcal{G}$-measurable;
2. $E|Y| < \infty$; and
3. for all $G \in \mathcal{G}$, $\int_G Y\, dP = \int_G X\, dP$, i.e., $E[Y 1_G] = E[X 1_G]$.

Any such Y is called *a version* of $E[X \mid \mathcal{G}]$.

Conditional expectation: Remarks on the definition

(1) "For every $G \in \mathcal{G}$" can be replaced by "For every $G \in \mathcal{P}$," where $\mathcal{P}$ is any π-system that generates $\mathcal{G}$.

Proof.

This is because

$$\mathcal{L} = \{G \in \mathcal{G} : E[Y1_G] = E[X1_G]\}$$

is a λ-system.

Why? By assumption, $\Omega \in \mathcal{L}$. If $G_1, G_2 \in \mathcal{L}$ with $G_1 \subset G_2$, then $G_2 \setminus G_1 \in \mathcal{L}$ as $1_{G_2 \setminus G_1} = 1_{G_2} - 1_{G_1}$. If $G_n \in \mathcal{L}$, $G_n \uparrow G$, then $1_{G_n}(\omega) \uparrow 1_G(\omega)$ for every $\omega \in \Omega$. So we can use DCT to conclude that $E[X1_{G_n}] \to E[X1_G]$, $E[Y1_{G_n}] \to E[Y1_G]$, and so $E[X1_G] = E[Y1_G]$, $G \in \mathcal{L}$.

Now, the π-λ theorem implies that $\mathcal{G} = \sigma(\mathcal{P}) \subset \mathcal{L}$. $\square$

Conditional expectation: Remarks on the definition

(2) $E[X \mid \mathcal{G}]$ is unique up to modifications on sets of measure 0. That is, if Y and Y' are both versions of $E[X \mid \mathcal{G}]$, then $P(Y = Y') = 1$.

Proof.

Take an $\epsilon > 0$. Take $A = \{Y - Y' \geq \epsilon\}$. Then $A \in \mathcal{G}$ and so $E[Y 1_A] = E[X 1_A] = E[Y' 1_A]$. Observe also that $(Y - Y') 1_A \geq \epsilon 1_A$. It follows that

$$0 = E[(Y - Y') 1_A] \geq \epsilon P(A),$$

and so $P(A) = 0$ for all $\epsilon > 0$. By taking intersection over countably many ϵ, $P(Y - Y' > 0) = 0$. By symmetry, $P(Y - Y' < 0) = 0$. □

Conditional expectation: Remarks on the definition

(3) Define

$$E[X \mid Z] = E[X \mid \sigma(Z)]$$
$$E[X \mid Z_1, \ldots, Z_n] = E[X \mid \sigma(Z_1, \ldots, Z_n)]$$

Here, $Z, Z_1, \ldots, Z_n$ are arbitrary r.v.'s. Sometimes, it is also written, for $B \in \mathcal{F}$,

$$P(B \mid \mathcal{G}) = E[1_B \mid \mathcal{G}],$$

but this is a bit confusing, as $P(B \mid A)$ (a number) is not the same as $E[1_B \mid 1_A]$ (a random variable).

Conditional expectation: Remarks on the definition

(4) Existence follows from the *Radon-Nikodym theorem*.

If μ, ν are σ-finite positive measures on $(\Omega, \mathcal{F})$, then we say that ν is *absolutely continuous* w.r.t. μ, denoted by $\nu \ll \mu$ if $\mu(A) = 0 \implies \nu(A) = 0$.

For practice, let's prove the following continuity characterization:
$\nu \ll \mu \iff (\forall \epsilon > 0)(\exists \delta > 0)(\forall A \in \mathcal{F})(\mu(A) < \delta \implies \nu(A) < \epsilon)$

Proof.

($\Longleftarrow$) If $\mu(A) = 0$, then $\nu(A) < \epsilon$ for all $\epsilon > 0$.
($\Longrightarrow$) Assume the negation:
$(\exists \epsilon > 0)(\forall \delta > 0)(\exists A_\delta \in \mathcal{F})(\mu(A_\delta) < \delta \;\&\; \nu(A_\delta) \geq \epsilon)$.

Take $A = \{A_{2^{-n}} \text{ i.o.}\}$.

Note that $\mu(A_{2^{-n}}) < 2^{-n}$ and $\nu(A_{2^{-n}}) \geq \epsilon$.

Then, by BC, $\mu(A) = 0$, but $\nu(A) \geq \limsup \nu(A_{2^{-n}}) \geq \epsilon$. □

Conditional expectation: Remarks on the definition

Aside. The measures μ and ν on $(\Omega, \mathcal{F})$ are *equivalent* if they have the same measure-zero sets: $\nu \ll \mu$ and $\mu \ll \nu$. For two equivalent probability measures, we cannot tell *with certainty* which one is used in a random experiment.

For example, suppose we have a fair coin and a unfair coin with heads probability $p \in (0, 1/2)$. If we toss each a finite number n times, the resulting two measures μ and ν are equivalent.

However, if we toss each infinitely many times, there exists a set A so that $\mu(A) = \nu(A^c) = 1$, i.e., the two measures are *orthogonal*, denoted $\mu \perp \nu$. Why is this true?

Conditional expectation: Remarks on the definition

Aside. The measures μ and ν on $(\Omega, \mathcal{F})$ are *equivalent* if they have the same measure-zero sets: $\nu \ll \mu$ and $\mu \ll \nu$. For two equivalent probability measures, we cannot tell *with certainty* which one is used in a random experiment.

For example, suppose we have a fair coin and a unfair coin with heads probability $p \in (0, 1/2)$. If we toss each a finite number n times, the resulting two measures μ and ν are equivalent.

However, if we toss each infinitely many times, there exists a set A so that $\mu(A) = \nu(A^c) = 1$, i.e., the two measures are *orthogonal*, denoted $\mu \perp \nu$. Why is this true?

SLLN! Let $X_k = 1_{\{k\text{'th toss H}\}}$. Let

$$A = \{\lim_{n\to\infty} (X_1 + \ldots + X_n)/n = 1/2\}.$$

Conditional expectation: Remarks on the definition

Theorem (Radon-Nikodym Theorem)

If $\nu \ll \mu$, then there exists an $\mathcal{F}$-measurable function $f \geq 0$ such that for every $A \in \mathcal{F}$, $\nu(A) = \int_A f\, d\mu$.

Observe that the reverse holds as well: if $\nu(A) = \int_A f\, d\mu$, then $\nu \ll \mu$.

The statement that $\nu(A) = \int_A f\, d\mu$ for every $A \in \mathcal{F}$ is often abbreviated as $d\nu = f\, d\mu$ or $f = d\nu/d\mu$, and f is called the *Radon-Nikodym derivative*.

Conditional expectation: Remarks on the definition

Existence of conditional expectation.

Suppose $X \geq 0$. Let $\mu = P$ and $\nu(A) = \int_A X\, dP$, viewed as measures on $(\Omega, \mathcal{G})$. Then $\nu \ll \mu$ so by RN, there exits a r.v. Y on $(\Omega, \mathcal{G})$ so that

$$\int_A X\, dP = \nu(A) = \int_A Y\, dP$$

for every $A \in \mathcal{G}$. For general X, split $X = X_+ - X_-$. $\square$

Note: the proof shows we can define $E[X \mid \mathcal{G}]$ for any $X \geq 0$.

Conditional expectation: Remarks on the definition

(5) Intuitively, $E[X \mid \mathcal{G}]$ is "the best guess of X based on information in $\mathcal{G}$." We will state this precisely later, but for now we make two observations.

If X is $\mathcal{G}$-measurable, then $E[X \mid \mathcal{G}] = X$.

If X is independent of $\mathcal{G}$, then $E[X \mid \mathcal{G}] = EX$, because for every $G \in \mathcal{G}$

$$E[X 1_G] = EX \cdot E 1_G = E[EX \cdot 1_G]$$

Conditional expectation: Properties

Assume that all conditional expectations below exist, i.e., all r.v.'s have finite expectation, and $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ are σ-algebras.

(1) Linearity: For $\alpha, \beta \in \mathbb{R}$,
$E[(\alpha X + \beta Y) \mid \mathcal{G}] = \alpha E[X \mid \mathcal{G}] + \beta E[Y \mid \mathcal{G}]$, a.s.

Proof.

Easy verification. □

(2) $E[E[X \mid \mathcal{G}]] = EX$.

Proof.

Apply the definition to $G = \Omega$. □

Conditional expectation: Properties

(3) Monotonicity: If $X \leq Y$ a.s., then $E[X \mid \mathcal{G}] \leq E[Y \mid \mathcal{G}]$ a.s.

Proof.

Let $G = \{E[X \mid \mathcal{G}] > E[Y \mid \mathcal{G}]\} \in \mathcal{G}$. Then

$$0 \leq E[(E[X \mid \mathcal{G}] - E[Y \mid \mathcal{G}])1_G] = E[(X - Y)1_G] \leq 0,$$

and so $P(G) = 0$. □

Conditional expectation: Properties

(4) Monotone convergence: If $0 \leq X_n \uparrow X$ (with finite expectation), then $E[X_n \mid \mathcal{G}] \uparrow E[X \mid \mathcal{G}]$ a.s.

Proof.

By (3), $E[X_n \mid \mathcal{G}] \uparrow Y$ a.s., for some $\mathcal{G}$-measurable r.v. Y. Take $G \in \mathcal{G}$. Then

$$E[E[X_n \mid \mathcal{G}]1_G] = E[X_n 1_G]$$

and, by MCT, the LHS converges to $E[Y1_G]$ while the RHS converges to $E[X1_G]$. So $E[X1_G] = E[Y1_G]$, Y has finite expectation, and, then, by definition $Y = E[X \mid \mathcal{G}]$ a.s. □

Conditional expectation: Properties

(5) Fatou: If $X_n \geq 0$, then $E[\liminf X_n \mid \mathcal{G}] \leq \liminf E[X_n \mid \mathcal{G}]$ a.s.

Proof.

Apply (4) to $X_n' = \inf\{X_n, X_{n+1}, \ldots\}$:
as $0 \leq X_n' \uparrow \liminf X_n$, $E[X_n' \mid \mathcal{G}] \uparrow E[\liminf X_n \mid \mathcal{G}]$, but by (3)
$E[X_n' \mid \mathcal{G}] \leq E[X_n \mid \mathcal{G}]$. □

Conditional expectation: Properties

(6) Dominant convergence: Assume $|X_n| \leq V$, and $EV < \infty$. If $X_n \to X$ a.s., then $E[X_n \mid \mathcal{G}] \to E[X \mid \mathcal{G}]$ a.s.

Proof.

WLOG, $X_n \geq 0$. Apply (5) to X_n and $V - X_n$. □

Conditional expectation: Properties

(7) Jensen's inequality: Assume that $\varphi : \mathbb{R} \to \mathbb{R}$ is convex and $E|\varphi(X)| < \infty$. Then $E[\varphi(X) \mid \mathcal{G}] \geq \varphi(E[X \mid \mathcal{G}])$ a.s.

For example, $E[|X| \mid \mathcal{G}] \geq |E[X \mid \mathcal{G}]|$, $E[X^2 \mid \mathcal{G}] \geq (E[X \mid \mathcal{G}])^2$.

Proof.

As φ is convex, we can write

$$\varphi(z) = \sup\{\alpha z + \beta : \alpha, \beta \in \mathbb{Q}, \alpha z + \beta \leq \varphi(z) \text{ for all } z\}.$$

If $\alpha z + \beta \leq \varphi(z)$ for all z, then $\alpha X + \beta \leq \varphi(X)$ and so $\alpha E[X \mid \mathcal{G}] + \beta \leq E[\varphi(X) \mid \mathcal{G}]$ a.s. Now take the supremum over α and β. □

Conditional expectation: Properties

(8) Tower property: If $\mathcal{H} \subset \mathcal{G}$, then

$$E[E[X \mid \mathcal{G}] \mid \mathcal{H}] = E[E[X \mid \mathcal{H}] \mid \mathcal{G}] = E[X \mid \mathcal{H}] \text{ a.s.}$$

Proof.

The second tower conditioning is clear, as $E[X \mid \mathcal{H}]$ is $\mathcal{G}$-measurable. For the first one, take $A \in \mathcal{H}$. We need to show that

$$E[E[X \mid \mathcal{G}]1_A] = E[X1_A],$$

which holds because $A \in \mathcal{G}$. □

Conditional expectation: Properties

(9) Taking out what is known: If Z is $\mathcal{G}$-measurable, and $E|XZ| < \infty$, then $E[XZ \mid \mathcal{G}] = ZE[X \mid \mathcal{G}]$ a.s.

Proof.

WLOG, $X \geq 0$. Take $Z = 1_A$ for some $A \in \mathcal{G}$. Then, for every $G \in \mathcal{G}$,

$$\begin{aligned} E[(X1_A)1_G] = E[X1_{A\cap G}] &= E[E[X \mid \mathcal{G}]1_{A\cap G}] \\ &= E[(E[X \mid \mathcal{G}]1_A)1_G], \end{aligned}$$

and so $E(X1_A \mid \mathcal{G}] = 1_A E(X \mid \mathcal{G}]$, i.e., the claim holds when Z is an indicator. Then, by (1), it holds when Z is simple, then, by (4), when Z is positive, and finally by (1) for arbitrary Z. □

Conditional expectation: Properties

(9) Discarding independent information: If $\mathcal{H}$ and $\sigma(\sigma(X) \cup \mathcal{G})$ are independent, then $E[X \mid \sigma(\mathcal{G} \cup \mathcal{H})] = E[X \mid \mathcal{G}]$ a.s.

Proof.

WLOG, $X \geq 0$. Let $Y = E[X \mid \mathcal{G}] \geq 0$.

Fix $G \in \mathcal{G}$, $H \in \mathcal{H}$. Then, by independence,

$$E[X1_{G \cap H}] = E[X1_G 1_H] = E[X1_G]P(H).$$

Applying the same reasoning to Y, we get

$$E[Y1_{G \cap H}] = E[Y1_G]P(H) = E[X1_G]P(H) = E[X1_{G \cap H}].$$

Now, $\{G \cap H : G \in \mathcal{G}, H \in \mathcal{H}\}$ is a π-system that generates $\sigma(\mathcal{G} \cup \mathcal{H})$, and Y is $\sigma(\mathcal{G} \cup \mathcal{H})$-measurable, and so $Y = E[X \mid \sigma(\mathcal{G} \cup \mathcal{H})]$. □

Conditional expectation: Properties

Example. Assume that X, Y are independent and equally distributed, with $P(X = \pm 1) = 1/2$. Let $Z = XY$, $\mathcal{G} = \sigma(Y)$ and $\mathcal{H} = \sigma(Z)$. Then X is independent of $\mathcal{G}$ (and also on $\mathcal{H}$), and $\mathcal{G}$ and $\mathcal{H}$ are independent. However, as $X = YZ$,

$$E[X \mid \sigma(\mathcal{G} \cup \mathcal{H})] = X$$

but

$$E[X \mid \mathcal{G}] = EX = 0.$$

Observe that $\mathcal{H}$ and $\sigma(\sigma(X) \cup \mathcal{G})$ are not independent.

Conditional expectation: Properties

Example. Assume $X_1, \dots, X_n$ are i.i.d., with $E|X_1| < \infty$, and let $S_n = X_1 + \cdots + X_n$. Compute $E[X_1 \mid S_n]$.

Conditional expectation: Properties

Example. Assume $X_1, \ldots, X_n$ are i.i.d., with $E|X_1| < \infty$, and let $S_n = X_1 + \cdots + X_n$. Compute $E[X_1 \mid S_n]$.

We have $E[X_1 \mid S_n] = E[X_i \mid S_n]$ for all i, because of symmetry. For example,

$$E[X_1 1_{\{S_n \in B\}}] = E[X_2 1_{\{S_n \in B\}}]$$

for all $B \in \mathcal{B}(\mathbb{R})$. So,

$$E[X_1 \mid S_n] = \frac{1}{n} E[S_n \mid S_n] = \frac{1}{n} S_n.$$

Conditional expectation: Special cases

(1) Assume $\Omega_1, \Omega_2, \ldots$ are disjoint measurable sets with nonzero probability such that $\Omega = \cup_i \Omega_i$. Let $\mathcal{G} = \sigma(\Omega_1, \Omega_2, \ldots)$. Then

$$E[X \mid \mathcal{G}] = \sum_i \frac{E[X 1_{\Omega_i}]}{P(\Omega_i)} \cdot 1_{\Omega_i}$$

Proof.

Denote the RHS by Y. Observe that that Y is $\mathcal{G}$-measurable and that $E|Y| \leq E|X|$. Notice that $\{\Omega, \emptyset, \Omega_j, j = 1, 2, \ldots\}$ is a π-system. For $A = \Omega_j$, clearly $E[X 1_A] = E[Y 1_A]$. By DCT, this is true for $A = \Omega$ as well. □

In particular,

$$E[1_B \mid 1_A] = P(B \mid A)\, 1_A + P(B \mid A^c)\, 1_{A^c}$$

Conditional expectation: Special cases

(2) Suppose X and Y have joint density $f(x,y)$, i.e.,

$$P((X,Y)\in B)=\int_B f(x,y)\,dxdy,\ \text{for every } B\in\mathcal{B}(\mathbb{R}^2)$$

Assume that $g:\mathbb{R}\to\mathbb{R}$ is Borel measurable and $E|g(X)|<\infty$. Then

$$E[g(X)\mid Y]=h(Y)$$

where

$$h(y)=\frac{\int_{\mathbb{R}} g(x)f(x,y)\,dx}{\int_{\mathbb{R}} f(x,y)\,dx}=\text{“}E[g(X)1_{Y\in[y,y+dy]}]/P(Y\in[y,y+dy])\text{”}.$$

Note that the denominator is the density $f_Y(y)$ of Y.
Using the “$0/0=0$” convention, the formula means that $h(y)=0$ when $f_Y(y)=0$. We call $f(x,y)/f_Y(y)$ the *conditional density* of X given $Y=y$.

Conditional expectation: Special cases

Proof.

WLOG, $g \geq 0$. Obviously, $h(Y)$ is $\sigma(Y)$-measurable. To verify the other two properties, take any $A \in \sigma(Y)$, that is, $A = \{Y \in B\}$ for some $B \in \mathcal{B}(\mathbb{R})$. Then

$$\begin{aligned}
E[h(Y)1_A] &= E[h(Y)1_{\{Y \in B\}}] \\
&= \int_B h(y) f_Y(y)\, dy = \int_B f_Y(y)\, dy \int_{\mathbb{R}} g(x) \frac{f(x,y)}{f_Y(y)}\, dx \\
&= \int_{\mathbb{R}^2} 1_{\{y \in B\}} g(x) f(x,y)\, dxdy \\
&= E[g(X)1_{\{Y \in B\}}] \\
&= E[g(X)1_A].
\end{aligned}$$

□

Conditional expectation: Special cases

Example. Let T_1, T_2 be independent Exp(1) random variables, and $S_1 = T_1$, $S_2 = T_1 + T_2$. Describe conditional distribution of S_1 given S_2.

$$
\begin{aligned}
f_{S_1,S_2}(s_1, s_2) &= f_{T_1,T_2}(s_1, s_2 - s_1) = e^{-s_1} e^{-(s_2 - s_1)} 1_{\{0 \le s_1 \le s_2\}} \\
&= e^{-s_2} 1_{\{0 \le s_1 \le s_2\}}
\end{aligned}
$$

and

$$f_{S_2}(s_2) = s_2 e^{-s_2} 1_{\{0 \le s_2\}}$$

The conditional density is the quotient, which equals

$$\frac{1}{s_2} 1_{\{0 \le s_1 \le s_2\}},$$

uniform on $[0, s_2]$.

Conditional expectation: Special cases

(3) Assume that X and Y are independent, and $\varphi : \mathbb{R}^2 \to \mathbb{R}$ Borel, with $E|\varphi(X,Y)| < \infty$.
Let $h(y) = E\varphi(X,y) = \int_{\mathbb{R}} \varphi(x,y)\,d\mu_X(x)$.
Then $E[\varphi(X,Y) \mid Y] = h(Y)$ (="$E_X[\varphi(X,Y)]$").

Proof.

WLOG, $\varphi \geq 0$. Again, take $A = \{Y \in B\}$, $B \in \mathcal{B}(\mathbb{R})$. Then

$$\begin{aligned}
E[h(Y)1_A] = E[h(Y)1_{\{Y\in B\}}] &= \int_{\mathbb{R}} h(y)1_{\{y\in B\}}\,d\mu_Y(y) \\
&= \int_{\mathbb{R}} 1_{\{y\in B\}}\,d\mu_Y(y) \int_{\mathbb{R}} \varphi(x,y)\,d\mu_X(x) \\
&= \int_{\mathbb{R}^2} 1_{\{y\in B\}}\varphi(x,y)\,d\mu_X(x)d\mu_Y(y) \\
&= \int_{\mathbb{R}^2} 1_{\{y\in B\}}\varphi(x,y)\,d\mu_{(X,Y)}(x,y) \\
&= E[\varphi(X,Y)1_A].
\end{aligned}$$

Regular conditional distribution

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(S, \mathcal{S})$ a set with a σ-algebra. Let $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ be a measurable map. Assume that $\mathcal{G} \subset \mathcal{F})$ is a σ-algebra. We say that $\mu : \Omega \times S \to [0, 1]$ is a *regular conditional distribution* of X given $\mathcal{G}$ if:

- for every $A \in \mathcal{S}$, $\mu(\cdot, A)$ is a version of $E[1_A \mid \mathcal{G}]$; and
- there exists a set $\Omega_0 \subset \Omega$ such that $P(\Omega_0) = 1$ and such that, for every $\omega \in \Omega_0$, $A \mapsto \mu(\omega, A)$ is a probability measure on $(S, \mathcal{S})$.

So μ is a *random* probability measure on $(S, \mathcal{S})$.

Note that we cannot just take $\mu(\cdot, A)$ to be *any* version of $E[1_A \mid \mathcal{G}]$ as we need countable additivity simultaneously for all pairwise disjoint countable collections of sets and for all $\omega \in \Omega_0$.

Regular conditional distribution

If we find a regular conditional distribution μ, then the standard argument (indicator $\to$ simple $\to$ positive $\to$ all) shows that for every measurable function $g : (S, \mathcal{S}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, such that $E|g(X)| < \infty$,

$$E[g(X) \mid \mathcal{G}] = \int_S g(x)\, \mu(\cdot, dx).$$

For example, if (X, Y) is a pair of r.v.'s with joint density f, $(S, \mathcal{S}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $\mathcal{G} = \sigma(Y)$, then r.c.d. is the (random) measure given by its density w.r.t. the Lebesgue measure:

$$\mu(\cdot, dx) = \frac{f(x, Y)}{f_Y(Y)}\, dx.$$

Regular conditional distribution

The existence of r.c.d. for all X is a property of the target space $(S, \mathcal{S})$. If $(S, \mathcal{S}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, then it always exists. More generally, it always exists when $\mathcal{S} = \sigma\{S_1, S_2, \ldots\}$ is generated by a countable collection of $S_i \in \mathcal{S}$; then we call $(S, \mathcal{S}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ a *Borel space*.

Any complete separable metric space (a.k.a. a Polish space) with a Borel σ-algebra is a Borel space, as its Borel σ-algebra is generated by the set of open balls of rational radii around points in a countable dense subset. Although there are counterexamples, for practical purposes r.c.d. always exists. While the proof of existence is not very difficult, it is not provided here.

Conditional expectation: L^2 theory

Theorem

Assume $EX^2 < \infty$. Then $E[X \mid \mathcal{G}]$ is the unique (up to a.s. equality) $\mathcal{G}$-measurable random variable Z that minimizes $E[(X - Z)^2]$.

Proof.

WLOG, $EZ^2 < \infty$, as otherwise $E[(X - Z)^2] = \infty$.
Let $Y = E[X \mid \mathcal{G}]$, so that $E[(X - Y) \mid \mathcal{G}] = 0$. Assume that W is a $\mathcal{G}$-measurable random variable with $EW^2 < \infty$. Then $(X - Y)W$ has finite expectation by Cauchy-Schwarz.
Moreover,

$$\begin{aligned} E[(X - Y)W] &= E[E[(X - Y)W \mid \mathcal{G}]] \\ &= E[W\, E[(X - Y) \mid \mathcal{G}]] = 0. \end{aligned}$$

Conditional expectation: L^2 theory

Proof, continued.

Therefore, as $Y - Z$ is a $\mathcal{G}$-measurable,

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

and is clearly minimized when $Z = Y$. □

Observe from the proof that $X - Y$ is orthogonal to the subspace of L^2 $\mathcal{G}$-measurable r.v.'s, so that $Y = E[X \mid \mathcal{G}]$ is the projection of X onto that space.