PT: INTEGRALS & EXPECTATIONS

Consider a discrete random variable X taking values $x_1, x_2, \dots$ with probability $p_i = \mathbb{P}(X = x_i)$. Suppose we perform n independent replications of a random experiment, denoting X_j the value in the j-th replication.

Put $n_i = \#\{j \leq n \mid X_j = x_i\}$
$=$ number of times the value x_i is observed during the first n experiments.

The average value after n trials is

$$\overline{X}_n := \frac{1}{n}\sum_{j=1}^{n} X_j = \frac{1}{n}\sum_{j=1}^{n}\sum_i x_i \, \mathbb{1}(X_j = x_i)$$

partition on the possible values of X_j

$$= \frac{1}{n}\sum_i x_i \underbrace{\sum_j \mathbb{1}(X_j = x_i)}_{= n_i}$$

$$= \sum_i x_i \boxed{\frac{n_i}{n}}$$

Frequency interpretation of probability gives
$$\frac{n_i}{n} \approx \mathbb{P}(X = x_i)$$

$$\approx \sum_i x_i \, \mathbb{P}(X = x_i)$$

You probably remember this expression from previous courses, and used it as a definition of the expected value of a discrete random variable X. Likewise, if X is AC with density f, the expected value of X is taken as $\int x f(x)\,dx$. However, this is more a computational rule, rather than a definition of expectation.

However, how would you make sense of the expression $\int x f(x)\,dx$ if X is defined over $(\Omega, \mathcal{F}, \mathbb{P})$, with Ω = space of continuous functions? [provided we can construct a suitable measure $\mathbb{P}$ on it]

$\Rightarrow$ We need a more general expression.

I. EXPECTED VALUE OF A RV

1 General Definition

- Start with indicators.

  Consider $(\Omega, \mathcal{F}, \mathbb{P})$ = probability space

  For $A \in \mathcal{F}$, put $X = \mathbb{1}_A = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases}$

  How to define $\mathbb{E}X$?

  Again, making use of frequency interpretation,

  $$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j = \frac{n_A}{n} \longrightarrow \mathbb{P}(A) \quad \text{as } n \to \infty$$

  same notation as on page 1.

  Thus, put $\boxed{\mathbb{E}X = \mathbb{P}(A) \quad \text{for } X = \mathbb{1}_A.}$

- Next, we want to generalize $\mathbb{E}X$ for simple RVs.

③

To do so, we are guided once again by the frequency interpretation of probability. In particular, we would like to keep the property of linearity:

$$n \text{ trials} \begin{cases} x_1, \dots, x_n \to \text{mean value } \bar{x} = \frac{1}{n}\sum x_i \\ y_1, \dots, y_n \to \text{mean value } \bar{y} = \frac{1}{n}\sum y_i \end{cases}$$

$\downarrow$ take the sum $\downarrow$

$x_1+y_1 \quad \dots \quad x_n+y_n$

$\downarrow$

mean value

$$\overline{x+y} = \frac{1}{n}\sum(x_i+y_i) = \frac{1}{n}\sum x_i + \frac{1}{n}\sum y_i = \bar{x}+\bar{y}$$

so it would be desirable to construct $\mathbb{E}$ such that $\mathbb{E}(X+Y) \approx \overline{x+y} = \bar{x}+\bar{y} \approx \mathbb{E}X + \mathbb{E}Y$

$\Rightarrow$ For a simple RV $\boxed{X = \sum_{k=1}^{n} \alpha_k \mathbb{1}_{A_k}}$, put

$$\boxed{\mathbb{E}X = \sum_{k=1}^{n} \alpha_k \mathbb{P}(A_k)}$$

not necessarily a partition of Ω

In particular, for a constant RV $X(\omega) = \alpha \mathbb{1}_\Omega$, this definition ensures that $\mathbb{E}X = \alpha$. Good.

Indeed,

$$\mathbb{E}X = \mathbb{E}\left(\sum_{k=1}^{n} \alpha_k \mathbb{1}_{A_k}\right) = \sum_{k=1}^{n} \alpha_k \mathbb{E}\mathbb{1}_{A_k} = \sum_{k=1}^{n} \alpha_k \mathbb{P}(A_k)$$

enforcing linearity for indicators

expected value of an indicator (page 2)

④

↳ Consequences: • Linearity of expectation for simple RVs follows

Indeed, take $X = \sum \alpha_i \mathbb{1}_{A_i}$, $Y = \sum \beta_j \mathbb{1}_{B_j}$ partitions of Ω

Then

$$\mathbb{E}(\alpha X + \beta Y) = \mathbb{E} \sum_{i,j} (\alpha\alpha_i + \beta\beta_j)\mathbb{1}_{A_i\cap B_j}$$

$\in \mathbb{R}$

$$= \sum_{i,j} (\alpha\alpha_i + \beta\beta_j)\mathbb{P}(A_i\cap B_j)$$

$$= \alpha \sum_i \alpha_i \underbrace{\boxed{\sum_j \mathbb{P}(A_i\cap B_j)}}_{= \mathbb{P}(A_i)} + \beta \sum_j \beta_j \underbrace{\boxed{\sum_i \mathbb{P}(A_i\cap B_j)}}_{= \mathbb{P}(B_j)}$$

$$= \alpha \sum_i \alpha_i \mathbb{P}(A_i) + \beta \sum_j \beta_j \mathbb{P}(B_j)$$

$$= \alpha \mathbb{E}X + \beta \mathbb{E}Y$$

- For a simple RV $X \geq 0$, we have $\mathbb{E}X \geq 0$.
- Monotonicity follows as well: if $X \leq Y$, for X and Y simple RVs, then $Y - X \geq 0$, and so

$$\mathbb{E}(Y-X) \geq 0$$

$$\overset{\text{(linearity)}}{\Rightarrow} \mathbb{E}Y - \mathbb{E}X \geq 0$$

$$\mathbb{E}Y \geq \mathbb{E}X.$$

Summarizing, for simple random variables $X = \sum_{k=1}^{n} \alpha_k \mathbb{1}_{A_k}$, defining $\mathbb{E}X = \sum_{k=1}^{n} \alpha_k \mathbb{P}(A_k)$ yields

- Linearity $\mathbb{E}(\alpha X + \beta Y) = \alpha\mathbb{E}X + \beta\mathbb{E}Y$
- Constants are expectations themselves $\mathbb{E}X = c$ for $X \equiv c$
- Monotonicity: $X \leq Y \Rightarrow \mathbb{E}X \leq \mathbb{E}Y$.

By the way, this definition is consistent. If

$$X = \sum_{k=1}^{n} \alpha_k 1\!\!1_{A_k} = \sum_{k=1}^{n'} \alpha'_k 1\!\!1_{A'_k}, \text{ then } \sum \alpha_k \mathbb{P}(A_k) = \sum \alpha'_k \mathbb{P}(A'_k).$$

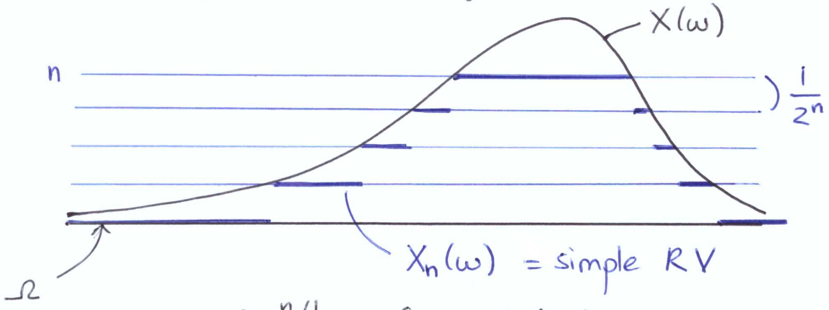
- Next, we consider general positive random variables $X \geq 0$. We know that there is a sequence of simple RVs $\{X_n\}$ such that

$$\forall \omega \in \Omega \quad X_n(\omega) \uparrow X(\omega) \text{ as } n \to \infty,$$

and this sequence can be constructed explicitely.

Set $A_{n,k} := \left\{ \omega \;\middle|\; \frac{k}{2^n} \leq X(\omega) < \frac{k+1}{2^n} \right\}$, $k = 0, 1, .., n2^n - 1$

$B_n := \{ \omega \mid X(\omega) \geq n \}$



Put $X_n(\omega) = \begin{cases} 2^n/k & \text{for } \omega \in A_{n,k} \\ n & \text{for } \omega \in B_n \end{cases}$

Monotonicity of $\{X_n\}$ follows directly since

$$A_{n,k} = A_{n+1,2k} \cup A_{n+1,2k+1}, \quad k < n2^n$$

X_n and X_{n+1} take the same value on this set

X_{n+1} is larger than X_n on this set (by $\frac{1}{2^{n+1}}$)

Thus $X_n \leq X_{n+1}$.

Moreover, for a fixed ω,

$$X_n(\omega) \leq X(\omega)$$

for $X(\omega) < n$ $\rightarrow$ $X_n(\omega) \geq X(\omega) - 2^{-n}$

$$0 \leq X(\omega) - X_n(\omega) \leq 2^{-n}$$

Thus $\boxed{X_n(\omega) \uparrow X(\omega) \text{ as } n \to \infty.}$

True for all ω such that $X(\omega) < \infty$, but it work for those ω's for which $X(\omega) = \infty$ as well!

$\Rightarrow$ Consequence: Since $\{X_n\}$ is a sequence of simple RVs such that $X_1 \leq X_2 \leq X_3 \leq \dots$, monotonicity implies that $\underset{\substack{\rotatebox{90}{$\leq$} \\ 0}}{\mathbb{E}X_1} \leq \mathbb{E}X_2 \leq \mathbb{E}X_3 \leq \dots$

$\Rightarrow$ the limit of this sequence exists, always! (but can be infinite).

Thus, $\boxed{\text{for a general } X \geq 0, \text{ we put } \mathbb{E}X := \lim_{n \to \infty} \mathbb{E}X_n}$

- Important remark: this is a consistent definition: the value of the limit does not depend on the choice of the sequence $\{X_n\}$. We only need that $X_n \uparrow X$.

Let's prove this.

⊗ Let $\{X_n\}$ and $\{\tilde{X}_n\}$ be two sequences of simple RVs such that $\forall \omega \in \Omega$ $X_n(\omega) \uparrow X(\omega)$, and $\tilde{X}_n(\omega) \uparrow X(\omega)$.

We want to show that $\lim_{n \to \infty} \mathbb{E} X_n = \lim_{n \to \infty} \mathbb{E} \tilde{X}_n$.

⊗ It suffices to prove that $\mathbb{E}\,\tilde{X}_k \leq \lim_{n\to\infty} \mathbb{E}\,X_n$ for any k, since this inequality implies that $\lim_{k\to\infty} \mathbb{E}\,\tilde{X}_k \leq \lim_{n\to\infty} \mathbb{E}\,X_n$. The inequality in the reverse direction follows by reversing the roles of X_n and $\tilde{X}_n$. The two limits must then coincide.

⊗ Put $A_n := \{\omega \mid X_n(\omega) \geq \tilde{X}_k(\omega) - \varepsilon\}$, for some fixed value of k and $\varepsilon > 0$. By definition of A_n, we have that $X_n \geq (\tilde{X}_k - \varepsilon)\, \mathbb{1}_{A_n}$

$$
\begin{aligned}
\mathbb{E}\,X_n &\geq \mathbb{E}\{(\tilde{X}_k - \varepsilon)\,\mathbb{1}_{A_n}\} && \text{monotonicity for simple RVs}\\
&= \mathbb{E}\,\tilde{X}_k \mathbb{1}_{A_n} - \varepsilon\,\mathbb{E}\,\mathbb{1}_{A_n} && \text{linearity}\\
&= \mathbb{E}\,\tilde{X}_k (1 - \mathbb{1}_{A_n^c}) - \varepsilon \underbrace{\mathbb{P}(A_n)}_{\leq 1}\\
&\geq \mathbb{E}\,\tilde{X}_k - \mathbb{E}\,\tilde{X}_k \mathbb{1}_{A_n^c} - \varepsilon\\
&\geq \mathbb{E}\,\tilde{X}_k - \Big[\max_{\omega\in\Omega} \tilde{X}_k(\omega)\Big]\,\mathbb{P}(A_n^c) - \varepsilon
\end{aligned}
$$

$\tilde{X}_k$ is a simple RV, so this value is $< \infty$.

Provided we show that $\mathbb{P}(A_n^c) \to 0$ as $n \to \infty$, we established that $\forall \varepsilon > 0$, $\lim_{n\to\infty} \mathbb{E}\,X_n \geq \mathbb{E}\,\tilde{X}_k - \varepsilon$

Since ε is arbitrary, we must have $\lim_{n\to\infty} \mathbb{E}\,X_n \geq \mathbb{E}\,\tilde{X}_k$.

⊗ Proof that $\mathbb{P}(A_n^c) \to 0$ as $n \to \infty$.
Since $X_{n+1} \geq X_n$, we have $A_n \subset A_{n+1}$. Moreover, since $X_n \uparrow X \geq \tilde{X}_k > \tilde{X}_k - \varepsilon$, one has $A_n \uparrow \Omega$.
Then $\mathbb{P}(A_n) \to 1$ (page 13 Chp "Solid Foundations") & thus $\mathbb{P}(A_n^c) \to 0$ ∎

• For a general (not necessarily positive) RV X, write $X = X^+ - X^-$, where

$$X^+ = \max(0, X) \geq 0$$
$$X^- = -\min(0, X) \geq 0$$

Note that in this notation, $|X| = X^+ + X^-$.

A random variable X is called INTEGRABLE if $\mathbb{E}|X| < \infty$, that if both $\mathbb{E}\,X^+$ and $\mathbb{E}\,X^-$ are finite. Since both X^+ and X^- are nonnegative, their expectation is well defined. Thus, for an integrable random variable X, put $\mathbb{E}\,X := \mathbb{E}X^+ - \mathbb{E}X^-$.

If one of the $\mathbb{E}\,X^{+/-}$ is infinite, we can still make use of this definition, which will be $\pm\infty$ depending on which $\mathbb{E}\,X^{+/-}$ is infinite.

However, if both $\mathbb{E}\,X^{+/-} = \infty$, then $\mathbb{E}\,X$ is undefined. (what is $\infty - \infty$?)

This idea of approximating a function by piecewise constant functions should look very familiar (remember the Riemann integral ?). In fact, the above construction is nothing else than the Lebesgue integral of X with respect to the probability measure $\mathbb{P}$ defined on $(\Omega, \mathcal{F})$.

X = simple RV on $[0,1]$

$\mathbb{E}X$ = area  (with respect to the Lebesgue measure)

⇒ Notation of $\mathbb{E}$ using integrals:

$$\mathbb{E}X = \int_\Omega X(\omega)\,\mathbb{P}(d\omega) = \int_\Omega X(\omega)\, d\,\mathbb{P}(\omega)$$

⇒ $\mathbb{E}X$ inherits all properties of Lebesgue integrals, including: 9

- Monotonicity: if $X \leq Y$ and $\mathbb{E}Y < \infty$, then $\mathbb{E}X \leq \mathbb{E}Y$

First consider non-negative X, Y, and approximate them using simple RVs $X_n \uparrow X$ and $Y_n \uparrow Y$. Then $\mathbb{E}X_n \leq \mathbb{E}Y_n$ follows from $X_n \leq Y_n$. The result follows by letting $n \to \infty$.

Next, drop the non-negativity assumption by considering separately X^+, Y^+ and X^-, Y^-.

- linearity: if both X and Y are integrable, then for $a, b \in \mathbb{R}$, $\mathbb{E}(aX + bY) = a\,\mathbb{E}X + b\,\mathbb{E}Y$.

First, note that $|aX + bY| \leq |a||X| + |b||Y|$, so that $\mathbb{E}|aX + bY| < \infty$ follows from $\mathbb{E}|X| < \infty$ and $\mathbb{E}|Y| < \infty$. The random variable $aX + bY$ is thus integrable.

Next, proceed as before.

Remarks
- For X integrable, we have that $|\mathbb{E}X| \leq \mathbb{E}|X|$.

Indeed, since by definition $\mathbb{E}X = \mathbb{E}X^+ - \mathbb{E}X^-$, we have that

$$
\begin{aligned}
|\mathbb{E}X| &\leq |\mathbb{E}X^+| + |\mathbb{E}X^-| \quad \text{since } X^{+/-} \geq 0 \\
&= \mathbb{E}X^+ + \mathbb{E}X^- \quad \text{linearity} \\
&= \mathbb{E}(X^+ + X^-) \\
&= \mathbb{E}|X|
\end{aligned}
$$

- For $\mathbb{C}$-valued RVs, expectations are defined component-wise

For $Z = X + iY$, $X, Y \in \mathbb{R}$, put $\mathbb{E}Z := \mathbb{E}X + i\,\mathbb{E}Y$ 10

Same for random vectors, expectation is defined component wise.

- The integral $\mathbb{E}X = \int_\Omega X(\omega)\,\mathbb{P}(d\omega)$ make sense when integrating functions X defined on more general spaces Ω than $\mathbb{R}$. For example, one may consider the space of continuous functions.

- A set $A \in \mathcal{F}$ is called a null set (with respect to $\mathbb{P}$) if $\mathbb{P}(A) = 0$. We say that two random variables X and Y are equal almost surely (a.s.), and we write $X = Y$ a.s. if $\{\omega \in \Omega \mid X(\omega) \neq Y(\omega)\}$ is a null set.

By construction of Lebesgue integral, it turns out that if $X = Y$ a.s, then X is integrable if and only if Y is integrable, and $\int X(\omega)\,\mathbb{P}(d\omega) = \int Y(\omega)\,\mathbb{P}(d\omega)$

X and Y have the same expected value.

And this holds true as well for inequalities. look: if $X \leq Y$ a.s. then $\mathbb{E}X \leq \mathbb{E}Y$

$$\left(\int X(\omega)\,\mathbb{P}(d\omega) \leq \int Y(\omega)\,\mathbb{P}(d\omega)\right)$$

(i) Show that if $X \geq 0$ a.s and $\mathbb{E}X = 0$, then $X = 0$ a.s.

(ii) Show that if $\mathbb{E}X < \infty$, then $X < \infty$ a.s.

If instead of considering $(\Omega, \mathcal{F}, \mathbb{P})$ and $X : \Omega \to \mathbb{R}$ with induced measure P_X, we take $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$ and set $X(x) = x$, using x in place of the usual ω, we see that $\int X(\omega)\, \mathbb{P}(d\omega) = \int x \, P_X(dx)$

More generally, if h = <u>measurable</u> function, then $Y := h(X)$ has expected value given by

$$\mathbb{E}Y = \mathbb{E}\, h(X) = \int_\Omega h(X(\omega))\, \mathbb{P}(d\omega)$$
$$= \int_{\mathbb{R}} h(x)\, P_X(dx)$$

as long as at least one of these two integrals make sense.

This property is know as "Théorème du Transfert", in French.
To prove this theorem, the strategy is to start with indicator functions $h = \mathbb{1}_B$, $B \in \mathcal{B}(\mathbb{R})$, then simple functions, then positive measurable functions (need monotone convergence theorem), and finally measurable functions. We will not go through details. Good.

<u>Consequences</u>: • For a discrete random variable X, the law of X is given by the discrete measure $P_X = \sum_x \mathbb{P}(X=x)\, \delta_x$

X is integrable if and only if $\mathbb{E}|X| = \sum_x |x|\, \mathbb{P}(X=x) < \infty$.

In this case $\mathbb{E}X = \sum_x x\, \mathbb{P}(X=x)$, and we recover the formula on page 1. Moreover, if h is measurable, then

$h(X)$ is a discrete random variable; it is integrable if and only if $\sum_x |h(x)|\, \mathbb{P}(X=x) < \infty$. Its expected value is then $\mathbb{E}\left[h(X)\right] = \sum_x h(x)\, \mathbb{P}(X=x)$.

• If X is AC with density f, the law of X satisfies

$$P_X(B) = \int_B f(x)\, dx \quad \forall B \in \mathcal{B}(\mathbb{R}).$$

To be more precise, AC ≡ with respect to the Lebesgue measure.

A consequence of the <u>RADON-NIKODYM THEOREM</u> (see below) yields $\mathbb{E}X = \int x f(x)\, dx$, provided X is integrable, that is $\mathbb{E}|X| = \int |x| f(x)\, dx < \infty$.

Likewise, if $h : \mathbb{R} \to \mathbb{R}$ is measurable, $h(X)$ is integrable if $\int_{\mathbb{R}} |h(x)|\, f(x)\, dx < +\infty$, and its expectation is $\mathbb{E}\left[h(X)\right] = \int_{\mathbb{R}} h(x) f(x)\, dx$.

$h(X)$ is not necessarily AC!

2

Let X = AC random variable
h = continuous function.
Construct X and h such that the random variable $h(X)$ is non-degenerate and discrete.

<u>Remark</u>: the Radon-Nikodym theorem. (RN)
Let $(\Omega, \mathcal{F})$ be a measurable space, endowed with two measures $\mathbb{P}$ and $\mathbb{Q}$. We say that $\mathbb{P}$ is absolutely continuous (AC

with respect to $\mathbb{Q}$ if $\forall A \in \mathcal{F}$, holds (13)

$$\mathbb{Q}(A) = 0 \Rightarrow \mathbb{P}(A) = 0,$$

and we write $\mathbb{P} \ll \mathbb{Q}$

A null set for $\mathbb{Q}$ is a null set for $\mathbb{P}$.

→ We have the following result (Radon-Nikodym)

If $\mathbb{P} \ll \mathbb{Q}$, then there exists a measurable function $f : (\Omega, \mathcal{F}) \to [0, +\infty)$ such that $\forall A \in \mathcal{F}$,

$$\mathbb{P}(A) = \int_A f \, d\mathbb{Q}$$

The function f is called the Radon-Nikodym density of $\mathbb{P}$ with respect to $\mathbb{Q}$. It is usually denoted $f = \frac{d\mathbb{P}}{d\mathbb{Q}}$.

Moreover, if h is measurable, we have

$$\int_\Omega h \, d\mathbb{P} = \int_\Omega h f \, d\mathbb{Q}$$

(as long as at least one of the two integrals make sense)

Not all distributions P_X are Absolutely Continuous with respect to the Lebesgue measure λ [defined such that $\lambda([a,b]) = b-a$ = length of the interval $[a,b]$, and extended to all Borel sets]

→ No discrete distribution is AC with respect to λ since $\mathbb{P}(X=x) > 0$ while $\lambda(\{x\}) = 0$.

→ Those that are AC w.r.t. λ are those with a density; that is the ones we encountered before.

Csq of RN theorem : $\left.\begin{array}{l} \mathbb{E}h(X) = \int h(x) P_X(dx) \\ P_X(B) = \int_B f(x)\lambda(dx) \end{array}\right\} \Rightarrow \mathbb{E}h(X) = \int h(x) f(x) \lambda(dx).$

(14)

Next, we give an alternative expression of the expected value of a non-negative random variable.

Theorem. Let $X \geq 0$ with distribution function F_X.

Then $\mathbb{E}X = \int_0^{+\infty} (1 - F_X(x)) \, dx$

$= \sum_{n \geq 1} n \, \mathbb{P}(X=n) = \sum_{n\geq 1} \mathbb{P}(X \geq n)$, if in addition X is integer valued.

proof: For $X \geq 0$, one has $X = \int_0^X dx = \int_0^{+\infty} \mathbb{1}(X > x) \, dx$

Thus $\mathbb{E}X = \mathbb{E}\int_0^{+\infty} \mathbb{1}(X>x)\,dx$

exchanging the order of integration (Tonelli)

$= \int_0^{+\infty} \mathbb{E}\,\mathbb{1}(X>x)\,dx$

$= \int_0^{+\infty} \mathbb{P}(X>x)\,dx \longrightarrow \overset{\text{discrete case}}{=} \sum_{k=1}^{+\infty} \int_{k-1}^{k} \mathbb{P}(X>x)\,dx$

$= \int_0^{+\infty} (1 - F_X(x))\,dx$

$= \sum_{k\geq 1} \int_{k-1}^{k} \mathbb{P}(X \geq k)\,dx$

$= \sum_{k \geq 1} \mathbb{P}(X \geq k) \underbrace{\int_{k-1}^{k} dx}_{=1}$

$= \sum_{k\geq 1} \mathbb{P}(X \geq k)$ ∎

15

Theorem. If • X_1 and X_2 are independent RVs
• g_1 and g_2 such that $g_i(X_i)$ is integrable,
Then $\mathbb{E}[g_1(X_1)\, g_2(X_2)] = \mathbb{E}[g_1(X_1)]\, \mathbb{E}[g_2(X_2)]$

Sketch of proof: • First, consider $g_i(x) = \mathbb{1}(x \in B_i)$ ← $\in \mathcal{B}(\mathbb{R})$.

Then
$$\begin{aligned}
\mathbb{E}[g_1(X_1)\, g_2(X_2)] &= \mathbb{E}[\mathbb{1}(X_1 \in B_1,\ X_2 \in B_2)] \\
&= \mathbb{P}(X_1 \in B_1,\ X_2 \in B_2) \\
&\overset{\text{independence}}{=} \mathbb{P}(X_1 \in B_1)\, \mathbb{P}(X_2 \in B_2) \\
&= \mathbb{E}\,\mathbb{1}(X_1 \in B_1)\ \mathbb{E}\,\mathbb{1}(X_2 \in B_2) \\
&= \mathbb{E}\, g_1(X_1)\ \mathbb{E}\, g_2(X_2).
\end{aligned}$$

• Then, consider simple functions
• Use these to approximate general functions. ∎

2 Moments & Spaces $\mathcal{L}^p$.

Moments are special cases of $\mathbb{E}\, h(X)$, with $h(x) = x^p$, $p \geq 1$.

The p-th moment of X is $\mathbb{E}\, X^p = \int x^p \, dF_X(x)$

The expected value of X is the first moment.

provided it exists! ⇒ integrability condition apply!

→ For a RV X and $p \geq 1$, let $\|X\|_p = \left(\mathbb{E}|X|^p\right)^{1/p}$.

A RV with $\|X\|_1 < \infty$ is called INTEGRABLE

$\|X\|_2 < \infty$ —"— SQUARE INTEGRABLE.

16

A random variable is called BOUNDED if there exists $K \in \mathbb{R}$ such that $|X| \leq K$ a.s.; the quantity $\|X\|_\infty$ is by definition the smallest such K.

→ The spaces $\mathcal{L}^p := \{ X : \Omega \to \mathbb{R} \mid \|X\|_p < \infty \}$ play a central role in functional analysis.

On $\mathcal{L}^p$, $\|\cdot\|_p$ is almost a norm (it is not a norm because $\|X\|_p = 0$ implies $X = 0$ a.s. and not $X(\omega) = 0$ for all ω)

• $\mathcal{L}^p$ is almost a BANACH SPACE.
• $\mathcal{L}^2$ is almost a HILBERT SPACE.

Define $\langle X, Y \rangle_{\mathcal{L}^2} = \mathbb{E}(XY)$, for $X, Y \in \mathcal{L}^2$

Remark: $X, Y \in \mathcal{L}^2 \Rightarrow \mathbb{E}(XY) < \infty$. Look:

$0 \leq (X \pm Y)^2 = X^2 + Y^2 \pm 2XY$

$\Rightarrow |XY| \leq \frac{1}{2}(X^2 + Y^2)$

Take $\mathbb{E}(\dots)$.

Result: $\langle \cdot, \cdot \rangle_{\mathcal{L}^2}$ is almost an inner product on $\mathcal{L}^2$.

(i) $\langle X+Y, Z \rangle_{\mathcal{L}^2} = \langle X, Z \rangle_{\mathcal{L}^2} + \langle Y, Z \rangle_{\mathcal{L}^2}$

(ii) $\langle \lambda X, Z \rangle_{\mathcal{L}^2} = \lambda \langle X, Z \rangle_{\mathcal{L}^2}$, $\lambda \in \mathbb{R}$

(iii) $\langle X, Z \rangle_{\mathcal{L}^2} = \langle Z, X \rangle_{\mathcal{L}^2}$

(iv) $\langle X, X \rangle_{\mathcal{L}^2} \geq 0$

However, $\langle X, X \rangle_{\mathcal{L}^2} = 0 \Rightarrow X = 0$ a.s.

In fact, we can modify the definition of $\mathcal{L}^2$ and regard two RVs in $\mathcal{L}^2$ as equal if they are equal $\mathbb{P}$-a.s, to make it a Hilbert space.

The induced norm is $\|X\|_2 = \sqrt{\langle X, X\rangle_{\mathcal{L}^2}} = \sqrt{\mathbb{E}X^2}$.

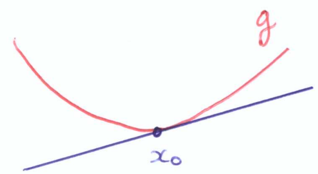
→ For $X \in \mathcal{L}^p$, the p-th <u>CENTRAL MOMENT</u> of X is $\mathbb{E}(X - \mathbb{E}X)^p$. The <u>VARIANCE</u> is the second central moment $\operatorname{Var} X = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \sigma_X^2$

σ_X = <u>STANDARD DEVIATION</u> (same scale as X)

Next, we present two elementary inequalities:

> <u>Theorem</u> (JENSEN INEQUALITY)
> Let $X \in \mathcal{L}^1$ and $g : \mathbb{R} \to \mathbb{R}$ a convex function.
> Then $g(\mathbb{E}X) \leqslant \mathbb{E}(g(X))$

<u>proof</u>



For a convex function g, $\forall x, x_0 \in \mathbb{R}$,

$$g(x) \geqslant g(x_0) + a(x - x_0),$$

where a is the gradient of g if g is differentiable. If not, then a is not unique.

Taking $x_0 = \mathbb{E}X$, $x = X$, $\quad g(X) \geqslant g(\mathbb{E}X) + a(X - \mathbb{E}X)$

(Take $\mathbb{E}(.)$ & use monotonicity + linearity.

$$\mathbb{E}[g(X)] \geqslant g(\mathbb{E}X) \quad \blacksquare$$

> <u>Theorem</u> (CHEBYSHEV / MARKOV)
> If $g : \mathbb{R} \to \mathbb{R}$ is a positive non-decreasing function, then for any RV X and $a \in \mathbb{R}$,
> $$\mathbb{P}(X > a) \leqslant \frac{\mathbb{E}[g(X)]}{g(a)}$$

<u>proof</u>: Since $g(a)\mathbb{1}(X \geqslant a) \leqslant g(X)$,

$$\mathbb{P}(X > a) = \mathbb{E}\,\mathbb{1}(X > a) \leqslant \mathbb{E}\left(\frac{g(X)}{g(a)}\right) = \frac{\mathbb{E}[g(X)]}{g(a)}$$

→ <u>Special cases</u>:

- $\mathbb{P}(|X| \geqslant a) \leqslant \dfrac{\mathbb{E}|X|^p}{a^p}$ for $p, a > 0$ $(X \in \mathcal{L}^p)$

  Take $g(x) = x^p$ and apply the inequality to $|X|$.

- $\mathbb{P}(|X - \mathbb{E}X| \geqslant a) \leqslant \dfrac{\operatorname{Var} X}{a^2}$ for $a > 0$

  Apply the inequality to $|X - \mathbb{E}X|$ and with $g(x) = x^2$

  ↳ In particular, we obtain the "3σ rule": the probability that a RV deviates from its mean by three σ is small: Take $a = 3\sigma$,

  $\mathbb{P}(|X - \mathbb{E}X| \geqslant 3\sigma) \leqslant 1/9$ (crude bound).

- $\mathbb{P}(X \geqslant a) \leqslant \dfrac{\mathbb{E}e^{tX}}{e^{ta}}$ for $t > 0$.

We present next further elementary results about the space $\mathcal{L}^p$.

Proposition

(i) $\mathcal{L}^p$ is a linear space: if $X, Y \in \mathcal{L}^p$, $\lambda \in \mathbb{R}$, then $X + Y \in \mathcal{L}^p$; $\lambda X \in \mathcal{L}^p$

(ii) If $X \in \mathcal{L}^p$, $1 \leq q \leq p$, then $\|X\|_q \leq \|X\|_p$; so that $\mathcal{L}^p \subset \mathcal{L}^q$ — Lyapunov inequality

(iii) HÖLDER'S INEQUALITY

Let $p, q \geqslant 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$.

If $X \in \mathcal{L}^p$ and $Y \in \mathcal{L}^q$, then $|\mathbb{E}XY| \leq \|X\|_p \|Y\|_q$

(iv) MINKOWSKI'S INEQUALITY

If $X, Y \in \mathcal{L}^p$, then $\|X+Y\|_p \leqslant \|X\|_p + \|Y\|_p$

proof (i) We implicitely used this result on page 16 when stating that $\mathcal{L}^p$ is a Banach space. It follows from $|x+y|^p \leqslant [2(\max(|x|,|y|))]^p$
$\leqslant 2^p(|x|^p + |y|^p)$.

(ii) This result implies that if the p-th moment is finite, then the q-th moment is finite. In particular, if the variance is finite, then the mean is finite. (look back at the definition of the p-th central moment on page 17)

It follows from Jensen's inequality with $g(x) = x^{p/q}$ for $x \geqslant 0$; which is convex for $p \geqslant q$. Putting $Y := |X|^q$,

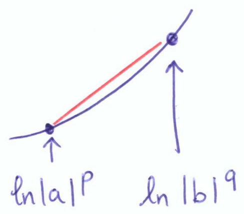
$$g(\mathbb{E}Y) = (\mathbb{E}Y)^{\frac{p}{q}} \leqslant \mathbb{E}(g(Y)) = \mathbb{E}\left(Y^{\frac{p}{q}}\right)$$

$$\left(\mathbb{E}|X|^q\right)^{\frac{1}{q}} \leqslant \left(\mathbb{E}|X|^p\right)^{1/p}.$$

(iii) Making use of the convexity of exp, we have that $\forall a, b$,

$$|ab| = \exp\Big[\underbrace{\tfrac{1}{p}}\ln|a|^p + \underbrace{\tfrac{1}{q}}\ln|b|^q\Big]$$

sum to 1



$$\leqslant \frac{1}{p}\exp\left[\ln|a|^p\right] + \frac{1}{q}\exp\left[\ln|b|^q\right]$$

$$= \frac{1}{p}|a|^p + \frac{1}{q}|b|^q.$$

Applying this to $a = \dfrac{|X|}{\|X\|_p}$; $b = \dfrac{|Y|}{\|Y\|_q}$, we obtain

$$\frac{|XY|}{\|X\|_p\|Y\|_q} \leqslant \frac{1}{p}\frac{|X|^p}{\|X\|_p^p} + \frac{1}{q}\frac{|Y|^q}{\|Y\|_q^q}$$

Taking $\mathbb{E}(.)$

$$\frac{\mathbb{E}|XY|}{\|X\|_p\|Y\|_q} \leqslant \frac{1}{p}\underbrace{\frac{\mathbb{E}|X|^p}{\|X\|_p^p}}_{=1} + \frac{1}{q}\underbrace{\frac{\mathbb{E}|Y|^q}{\|Y\|_q^q}}_{=1}$$

$$= \frac{1}{p} + \frac{1}{q} = 1$$

(iv). Let $q^{-1} = 1 - p^{-1}$ ($\Leftrightarrow pq^{-1} = p-1 \Leftrightarrow p = q(p-1)$)

- First, note that $|X+Y|^{p-1} \in \mathcal{L}^q$.
  Indeed, $\underset{\mathcal{L}^p}{\underset{\cap}{|X+Y|^p}} = |X+Y|^{p(q-1)}$. thus integrable.
- $\mathbb{E}|X+Y|^p \leqslant \mathbb{E}|X||X+Y|^{p-1} + \mathbb{E}|Y||X+Y|^{p-1}$

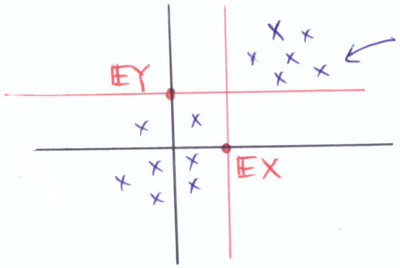
Hölder $\leqslant (\|X\|_p + \|Y\|_p)\, \| |X+Y|^{p-1} \|_q$

$$= (\|X\|_p + \|Y\|_p)\left(\mathbb{E}|X+Y|^p\right)^{\frac{1}{q}}$$

③ Covariance & Correlation

- For $X, Y \in \mathcal{L}^2$, the COVARIANCE of X and Y is defined as $\mathrm{Cov}(X,Y) = \mathbb{E}\left[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\right]$
$$= \mathbb{E}(XY) - \mathbb{E}X\, \mathbb{E}Y$$

Remember, for $X, Y \in \mathcal{L}^2$, we have that $XY \in \mathcal{L}^1$!



If X and Y tend to be 'large' together or 'small' together, the covariance of X and Y is positive.

- The CORRELATION between X and Y, provided $\mathrm{Var}\, X$, $\mathrm{Var}\, Y > 0$ is

$$\rho := \mathrm{Corr}(X,Y) := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}\, X \; \mathrm{Var}\, Y}} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

- Properties of the correlation coefficient

We made use on page 16 that for $X, Y \in \mathcal{L}^2$, $|XY| \leq \frac{1}{2}(X^2 + Y^2)$.

Applying this inequality with $X \to \frac{X}{\sqrt{\mathbb{E}X^2}}$

$Y \to \frac{Y}{\sqrt{\mathbb{E}Y^2}}$, we get

$$\left| \frac{XY}{\sqrt{\mathbb{E}X^2\, \mathbb{E}Y^2}} \right| \leq \frac{1}{2}\left(\frac{X^2}{\mathbb{E}X^2} + \frac{Y^2}{\mathbb{E}Y^2} \right)$$

Taking $\mathbb{E}(.)$, this leads to the famous...

CAUCHY–BUHYAKOVSKY INEQUALITY

If $X, Y \in \mathcal{L}^2$, then $XY \in \mathcal{L}^1$, and $\mathbb{E}|XY| \leq \sqrt{\mathbb{E}X^2\, \mathbb{E}Y^2}$

This is just a special case of Hölder's inequality, by the way. Take $p = q = 2$.

Moreover, since $|\mathbb{E}XY| \leq \mathbb{E}|XY|$, (Jensen)
replace X by $X - \mathbb{E}X$
Y by $Y - \mathbb{E}Y$; & it follows from CB inequality that $|\mathrm{Cov}(X,Y)| \leq \sqrt{\mathrm{Var}\, X\; \mathrm{Var}\, Y}$,
and thus that $|\mathrm{Corr}(X,Y)| \leq 1$

When do we have an equality here?

↳ Assume that $\mathrm{Corr}(X,Y) = +1$. Then, for the standardized RVs $X_1 = \frac{X - \mathbb{E}X}{\sqrt{\mathrm{Var}\, X}}$ and $Y_1 = \frac{Y - \mathbb{E}Y}{\sqrt{\mathrm{Var}\, Y}}$, we have

$$\mathbb{E}(X_1 - Y_1)^2 = \underbrace{\mathbb{E}X_1^2}_{=1} + \underbrace{\mathbb{E}Y_1^2}_{=1} - 2\underbrace{\mathbb{E}(X_1 Y_1)}_{= \mathrm{Corr}(X,Y) = 1} = 0$$

Likewise, when $\mathrm{Corr}(X,Y) = -1$, we have that $\mathbb{E}(X_1 + Y_1)^2 = 0$

Either way, we have a RV $Z := (X_1 \pm Y_1)^2 \geq 0$ with $\mathbb{E}Z = 0$. We know from on page 10 that this implies that $Z = 0$ a.s. We just showed that:

$$\mathrm{Corr}(X,Y) = \pm 1 \Leftrightarrow \mathbb{P}(X_1 \pm Y_1 = 0) = 1$$

Thus, with probability 1,

$$\frac{X - \mathbb{E}X}{\sqrt{\operatorname{Var} X}} \pm \frac{Y - \mathbb{E}Y}{\sqrt{\operatorname{Var} Y}} = 0$$

$$\Rightarrow Y = \underset{\substack{\uparrow \\ \text{Same sign as } \operatorname{Corr}(X,Y).}}{a} X + b$$

$\Rightarrow$ Correlation between X and Y is ± 1 where there is a perfect linear relationship between X and Y.

When $\operatorname{Corr}(X,Y) = 0$, we say that X and Y are <u>UNCORRELATED</u>

⚠ which is not the same as <u>INDEPENDENCE</u>.

If X and Y are independent, then $\mathbb{E}(XY) = \mathbb{E}X\,\mathbb{E}Y$, so that $\operatorname{Cov}(X,Y) = 0$ and X & Y are uncorrelated. But the converse does not hold in full generality.

Summarizing $\boxed{\text{Independence} \underset{\cancel{\Leftarrow}}{\Rightarrow} \text{Uncorrelation}}$

3

Provide examples of uncorrelated random variables that are not independent

<u>Remark</u>: When dealing with random vectors $X = (X_1, \dots, X_d)^t$, we use <u>COVARIANCE MATRICES</u>.

$$\underset{(d\times d)}{\Sigma} := \begin{bmatrix} \operatorname{Cov}(X_1, X_1) & \operatorname{Cov}(X_1, X_2) & \cdots \\ \operatorname{Cov}(X_2, X_1) & \operatorname{Cov}(X_2, X_2) & \cdots \\ \vdots & \vdots & \end{bmatrix} = \mathbb{E}\left[(X - \mathbb{E}X)(X-\mathbb{E}X)^t\right]$$

Two key properties: [P1] Σ is symmetric

[P2] Σ is positive semi-definite: $\forall x \in \mathbb{R}^d, \quad x^t \Sigma x \geq 0$

- [P1] is obvious
- To get [P2], Put $Y = X^t x \in \mathbb{R}$. Then

$$\begin{aligned}
0 \leq \operatorname{Var} Y &= \mathbb{E}(Y - \mathbb{E}Y)^2 \\
&= \mathbb{E}(X^t x - \mathbb{E}X^t x)^2 \\
&= \mathbb{E}\Big[\underbrace{(X - \mathbb{E}X)^t x}_{\in \mathbb{R}}\Big]^2 \\
&= \mathbb{E}\left[(X-\mathbb{E}X)^t x \,(X - \mathbb{E}X)^t x\right]^2 \\
&= \mathbb{E}\left[\{(X-\mathbb{E}X)^t x\}^t (X-\mathbb{E}X)^t x\right]^2 \\
&= x^t\, \mathbb{E}\left[(X - \mathbb{E}X)(X - \mathbb{E}X)^t\right] x \\
&= x^t \Sigma x
\end{aligned}$$

In fact, any $(d \times d)$ matrix satisfying [P1]+[P2] is the covariance matrix of some distribution on $\mathbb{R}^d$ (why?)

- <u>Geometrical Considerations.</u>

Consider the space of centered square integrable random variables $\mathcal{L}^2_0 := \{X \in \mathcal{L}^2 \mid \mathbb{E}X = 0\}$.

Let $X, Y \in \mathcal{L}^2_0$.

Then $\|X + Y\|^2_{\mathcal{L}^2} = \mathbb{E}(X+Y)^2$
$= \mathbb{E}X^2 + \mathbb{E}Y^2 + 2\,\mathbb{E}(XY)$

$$\Rightarrow \boxed{\|X+Y\|^2_{\mathcal{L}^2} = \|X\|^2_{\mathcal{L}^2} + \|Y\|^2_{\mathcal{L}^2} + 2\langle X, Y\rangle.}$$

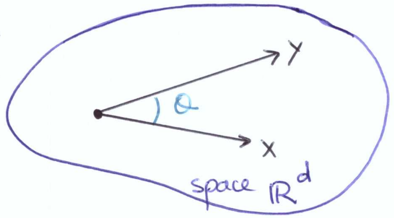
$$\operatorname{Var}(X+Y) = \operatorname{Var} X + \operatorname{Var} Y + 2\operatorname{Cov}(X,Y)$$

Compare the last written equality with what happens in $\mathbb{R}^d$, endowed with the usual metric $\langle \cdot,\cdot \rangle_d$ (s.t. $\forall x, y \in \mathbb{R}^d$, $\langle x, y \rangle_d = x^t y$).

$$\|x+y\|_d^2 = \langle x+y, x+y \rangle_d$$
$$= \langle x, x \rangle_d + \langle y, y \rangle_d + 2\langle x, y \rangle_d$$

$$\Rightarrow \boxed{\|x+y\|_d^2 = \|x\|_d^2 + \|y\|_d^2 + 2\langle x, y \rangle_d.}$$

Moreover, the cosine of the angle between x and y is $\cos\theta = \dfrac{\langle x, y \rangle_d}{\|x\|_d \|y\|_d}$



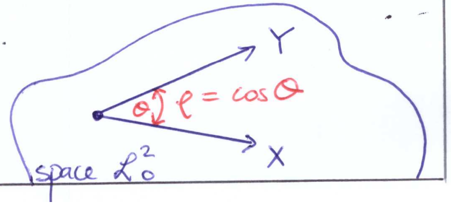
$\Rightarrow$ Therefore, for $X, Y \in \mathcal{L}_0^2$,

$$\rho = \mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}(XY)}{\sigma_X \sigma_Y}$$
$$= \frac{\langle X, Y \rangle_d}{\|X\|_{\mathcal{L}^2} \|Y\|_{\mathcal{L}^2}} = \cos\theta$$

since
$$\sigma_X = \sqrt{\mathrm{Var}\, X} = \sqrt{\mathbb{E}X^2} = \sqrt{\langle X, X \rangle_{\mathcal{L}^2}} = \|X\|_{\mathcal{L}^2}$$

$\Rightarrow$ The correlation coefficient represents the cosine of the angle between X and Y in the space $\mathcal{L}_0^2$.

In particular, when $\rho = 0$, X and Y are uncorrelated, $\theta = \frac{\pi}{2}$: X and Y are "perpendicular" and we write $X \perp Y$.



II_ CONDITIONAL EXPECTATION.

1) General definition.

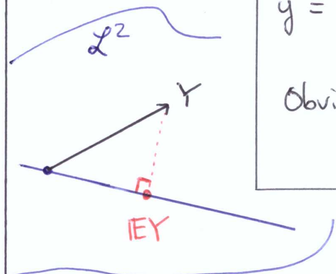
Motivation: goal is to estimate the value of a variable of interest, denoted Y. If no further information other than observed values of Y are provided, what is your best educated guess?

↳ 'best' must be explicitely defined. We place ourselves in the space $\mathcal{L}^2$ of square integrable RVs, endowed with the inner product $\langle X, Y \rangle_{\mathcal{L}^2} = \mathbb{E}(XY)$. The induced metric is $d^2(X,Y) = \mathbb{E}[(X-Y)^2]$.

- If we do not know anything about Y, you may want to estimate Y using a constant value $\hat{y}$ such that $\hat{y}$ minimizes the distance $d^2(Y, y)$:

$$\hat{y} = \underset{y}{\mathrm{argmin}}\ d^2(Y,y) = \underset{y}{\mathrm{argmin}}\ \mathbb{E}[(Y-y)^2]$$

Obviously, $\hat{y} = \mathbb{E}Y$.



3 Compute $\tilde{y} = \underset{y}{\mathrm{argmin}}\ \mathbb{E}|Y-y|$

- However, in a supervised learning context, we often know something about Y. For example, we may not know if a patient has some desease, but we might know the result of a medical test.

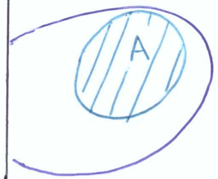
↳ Such extra information is commonly refferred to as a 'predictor'; a 'feature'; or a 'covariate'.

27

- Suppose the only available extra piece of information is the answer to a 'yes/no' question: we know that some event A occured (for example, blood pressure is higher than some threshold). What is your **best** educated guess of Y in this case.

↳ Proceed as before, and consider the minimization of $h(y) := \mathbb{E}\left[(Y-y)^2 \mathbb{1}_A\right]$

Restrict the minimization on a subset of Ω.



$$h(y) = \mathbb{E}\left[Y^2 \mathbb{1}_A\right] - 2y\,\mathbb{E}\left[Y \mathbb{1}_A\right] + y^2\,\mathbb{P}(A)$$

$$h'(y) = -2\,\mathbb{E}\left[Y\mathbb{1}_A\right] + 2\hat{y}\,\mathbb{P}(A) = 0$$

Gives $\hat{y} = \dfrac{\mathbb{E}[Y\mathbb{1}_A]}{\mathbb{P}(A)}$

ntuitively, responds to the ean value of Y fter discarding all runs of e experiment here A did not occur.

Notation: We use $\mathbb{E}(Y;A)$ to denote $\mathbb{E}[Y\mathbb{1}_A]$

- We can do the same for A^c instead of A. Our 'best' guess would be $\dfrac{\mathbb{E}[Y\mathbb{1}_{A^c}]}{\mathbb{P}(A^c)}$.

- In this simple situation, our predictor is the variable $X = \mathbb{1}_A$. Summarizing our findings, our best predictor of Y given X, that we denote $\hat{Y}$, is

$$\hat{Y}(\omega) = \begin{cases} \dfrac{\mathbb{E}[Y;A]}{\mathbb{P}(A)} & \text{if } \omega \in A \\ \dfrac{\mathbb{E}[Y;A^c]}{\mathbb{P}(A^c)} & \text{if } \omega \notin A \end{cases}$$

This object is a Random Variable!

28

⇒ Rewriting $\hat{Y}$ slightly differently,

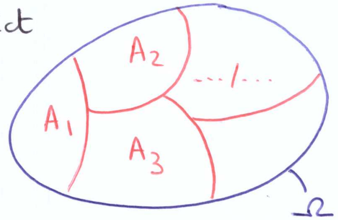
$$\hat{Y} = \begin{cases} \mathbb{E}[Y;A]/\mathbb{P}(A) & \text{with probability } \mathbb{P}(A) \\ \mathbb{E}[Y;A^c]/\mathbb{P}(A^c) & \text{— " —} \quad \mathbb{P}(A^c) \end{cases}$$

The mean value of $\hat{Y}$...

$$\dots \frac{\mathbb{E}[Y\mathbb{1}_A]}{\mathbb{P}(A)} \times \mathbb{P}(A) + \frac{\mathbb{E}[Y\mathbb{1}_{A^c}]}{\mathbb{P}(A^c)}\mathbb{P}(A^c) = \mathbb{E}(Y)$$

... equal to the expected value of Y. OK.

- Next, suppose that your predictor is a simple RV, $X = \sum_{i=1}^{n} x_i \mathbb{1}_{A_i}$, where $\{A_1, \dots, A_n\}$ is a partition of Ω, and all x_i are distinct



↳ If you observe $X = x_i$, then $\omega \in A_i$, and based on previous calculations, the 'best' forecast for Y is

$$\hat{Y}(\omega) := \frac{\mathbb{E}[Y;A_i]}{\mathbb{P}(A_i)} =: y_i, \quad \omega \in A_i$$

provided $\mathbb{P}(A_i) > 0$.

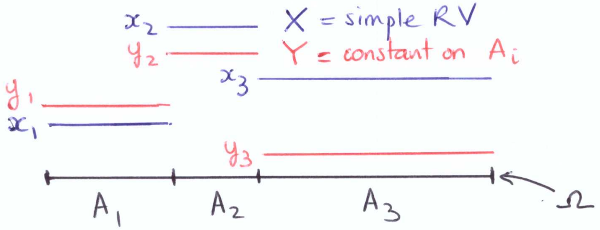
Thus

$$\hat{Y}(\omega) = \begin{cases} \dfrac{\mathbb{E}[Y;A_1]}{\mathbb{P}(A_1)} = y_1, & \omega \in A_1 \\ \vdots \\ \dfrac{\mathbb{E}[Y;A_n]}{\mathbb{P}(A_n)} = y_n, & \omega \in A_n \end{cases}$$

Rk: If $\mathbb{P}(A_i) = 0$, define $\hat{Y}$ as you wish on A_i ⇒ the resulting definition of $\hat{Y}$ is unique $\mathbb{P}$-almost-surely.

This object is a random variable!

Since $A_i = \{\omega \mid X(\omega) = x_i\}$, we introduce a function $\varphi(x)$ by putting $\varphi(x_i) = y_i$. We see that the new random variable $\hat{Y}$ is a function of X since in this notation, $\hat{Y} = \varphi(X)$.



Remarks:

- In view of the expression of $\hat{Y}$, $\hat{Y}$ represents the 'conditional mean of Y given X'. Instead of $\hat{Y}$, we may use the notation $\mathbb{E}(Y \mid X)$.

- The values of X do not matter when defining $\mathbb{E}(Y \mid X)$. What matters is the partition created by X. In other words, what matters is $\sigma(X)$. Therefore, we may use the notation $\mathbb{E}(Y \mid \sigma(X))$ in place of $\mathbb{E}(Y \mid X)$.

  [If you are interested in the average weight of inhabitants in a big city given their postcode; does the postcode itself actually matter, or the partition of the city it creates?]

- Y and $\hat{Y}$ have the same average value over A_i:

$$\mathbb{E}(\hat{Y}; A_i) = \mathbb{E}(\hat{Y}\, \mathbb{1}_{A_i}) = \mathbb{E}(y_i\, \mathbb{1}_{A_i}) = y_i\, \mathbb{P}(A_i) = \mathbb{E}(Y; A_i)$$

- Y and $\hat{Y}$ also have the same average over arbitrary union of A_i; that is over any element in $\sigma(X) = \{X^{-1}(B) \mid B \in \mathcal{B}(\mathbb{R})\}$.

Consider

$$X^{-1}(B) = \{X \in B\} = \bigcup_{k \mid x_k \in B} \{X = x_k\},$$

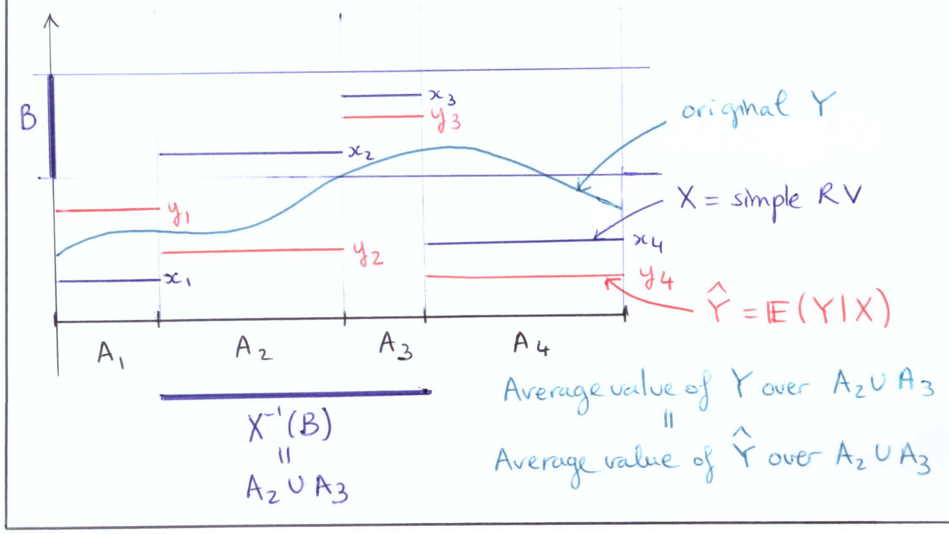
so that

$$\mathbb{1}_{\{X \in B\}} = \sum_{k \mid x_k \in B} \mathbb{1}_{A_k}.$$

Hence,

$$\begin{aligned}
\mathbb{E}(\hat{Y}; X \in B) &= \mathbb{E}(\hat{Y}\, \mathbb{1}_{\{X \in B\}}) \\
&= \mathbb{E}\Big(\hat{Y} \sum \mathbb{1}_{A_k}\Big) \\
&= \sum \mathbb{E}(\hat{Y}\, \mathbb{1}_{A_k}) \quad \text{page 29} \\
&= \sum \mathbb{E}(Y\, \mathbb{1}_{A_k}) \\
&= \mathbb{E}\Big(Y \sum \mathbb{1}_{A_k}\Big) \\
&= \mathbb{E}(Y\, \mathbb{1}_B) \\
&= \mathbb{E}(Y; X \in B), \text{ indeed.}
\end{aligned}$$

Picture =



In summary, we see that the conditional expectation $\hat{Y} = \mathbb{E}(Y|X)$ satisfies two fundamental properties. (31)

[CE.1] $\hat{Y}$ is a function of X: $\hat{Y} = \varphi(X)$

[CE.2] $\forall B \in \mathcal{B}(\mathbb{R})$, $\mathbb{E}[\hat{Y}; X \in B] = \mathbb{E}[Y; X \in B]$

$\{ \forall A \in \sigma(X)$, $\mathbb{E}[\hat{Y}; A] = \mathbb{E}[Y; A] \}$

<u>Remark</u>: These two conditions uniquely specify $\hat{Y}$ in the case of simple RVs. Indeed, with $A = A_i$ in [CE.2], we have $\mathbb{E}[\hat{Y}; A_i] = \mathbb{E}[Y; A_i]$.

Using [CE.1], $\hat{Y} = \varphi(X)$, so that

$$\begin{aligned} \mathbb{E}[\varphi(X) \mathbb{1}_{A_i}] &= \mathbb{E}[Y \mathbb{1}_{A_i}] \\ \parallel \quad & \\ \varphi(x_i)\, \mathbb{E}\, \mathbb{1}_{A_i} &= \varphi(x_i)\, \mathbb{P}(A_i), \end{aligned}$$

so that $\varphi(x_i) = \dfrac{\mathbb{E}[Y \mathbb{1}_{A_i}]}{\mathbb{P}(A_i)}$ = value of $\hat{Y}$ on A_i

We formally define the Conditional Expectation (CE) for a general X using [CE.1] and [CE.2]:

> **Theorem**: Let $Y \in \mathcal{L}^1$ and X be RVs defined on a common probability space.
> Then there exists a unique random variable satisfying [CE.1] and [CE.2], and is called the CE of Y given X; denoted $\mathbb{E}(Y/X)$. Moreover, $\mathbb{E}(Y|X) \in \mathcal{L}^1$.

Unique up to $\mathbb{P}$-null sets.

<u>proof</u>:

× <u>Uniqueness</u>: let $\hat{Y}'$ and $\hat{Y}''$ two RV in $\mathcal{L}^1$ such that $\forall A \in \sigma(X)$, $\mathbb{E}[\hat{Y}' \mathbb{1}_A] = \mathbb{E}[Y \mathbb{1}_A] = \mathbb{E}[\hat{Y}'' \mathbb{1}_A]$.

Take $A = \{\hat{Y}' > \hat{Y}''\} \in \sigma(X)$ since from [CE.1] both $\hat{Y}'$ and $\hat{Y}''$ are functions of X.

Then $\underbrace{\mathbb{E}[(\hat{Y}' - \hat{Y}'') \mathbb{1}_{\{\hat{Y}' > \hat{Y}''\}}]}_{\text{a positive random variable}} = 0$ (32)

$\Rightarrow (\hat{Y}' - \hat{Y}'') \mathbb{1}_{\{\hat{Y}' > \hat{Y}''\}} = 0$ a.s.

Hence $\hat{Y}' \leq \hat{Y}''$ a.s.

By symmetry, $\hat{Y}' \geq \hat{Y}''$ a.s, and we conclude that $\hat{Y}' = \hat{Y}''$ a.s.

× <u>Existence</u>: Suppose $Y \geq 0$.

Define a measure $\mathbb{Q}$ on $(\Omega, \sigma(X))$ by

$\forall A \in \sigma(X)$, $\mathbb{Q}(A) = \mathbb{E}[X \mathbb{1}_A]$

(not a proba measure).

Since $\mathbb{P}$ is a probability measure defined on $(\Omega, \sigma(X))$, we see that $\mathbb{Q} \ll \mathbb{P}$, by definition of $\mathbb{Q}$ $\left(\mathbb{Q}(A) = \int_A X(\omega)\, d\mathbb{P} \right)$.

$\Downarrow$

$\mathbb{P}(A) = 0 \Rightarrow \mathbb{Q}(A) = 0$

Radon-Nikodym theorem (page 13) ensures the existence of a measurable function $\tilde{X} : (\Omega, \sigma(X)) \longrightarrow \mathbb{R}$ such that

$$\mathbb{Q}(A) = \int_A \tilde{X}(\omega)\, d\mathbb{P} = \mathbb{E}[\tilde{X} \mathbb{1}_A]$$

Thus, $\tilde{X}$ is such that $\forall A \in \sigma(X)$, $\mathbb{E}[X \mathbb{1}_A] \overset{\parallel}{} \mathbb{E}[\tilde{X} \mathbb{1}_A]$

$\hookrightarrow \in \mathcal{L}^1$: take $A = \Omega$

and thus verifies [CE.2]

If Y is not ≥ 0, take $Y = Y^+ - Y^-$ ∎

33

Remark: Since $\mathbb{E}(Y|X)$ does not depend on the value of X but on the partition it creates, if φ is a one-to-one function, then $\mathbb{E}(Y|X) = \mathbb{E}(Y|\varphi(X))$.

Ex: $\mathbb{E}(Y|X) = \mathbb{E}(Y|X^3) = \mathbb{E}(Y|e^X)$ ...

Example = Let $X \sim \mathcal{P}(\lambda)$ $\quad \mathbb{P}(X=x) = \dfrac{\lambda^x}{x!} e^{-\lambda}$ $\quad x = 0, 1, \dots$

$Y \sim \mathcal{P}(\mu)$

$Z := X + Y \sim \mathcal{P}(\lambda+\mu)$

Show that $\mathbb{E}(X|Z) = \dfrac{\lambda}{\lambda+\mu} Z$

[CE.1] is obvious since $\dfrac{\lambda}{\lambda+\mu}$ is a function of Z.

[CE.2] $\mathbb{E}[X ; Z=k] = \mathbb{E}[X \, \mathbb{1}(X+Y=k)]$

$$= \sum_{i,j\geq 0} i \, \underbrace{\mathbb{1}(i+j=k)}_{j=k-i\geq 0} \, \mathbb{P}(X=i, Y=j)$$

$$= \sum_{i=0}^{k} i \, \mathbb{P}(X=i, Y=k-i)$$

$$= \sum_{i=0}^{k} i \, \frac{\lambda^i}{i!} e^{-\lambda} \frac{\mu^{k-i}}{(k-i)!} e^{-\mu}$$

$$= e^{-(\lambda+\mu)} \sum_{i=\cancel{0}\,1}^{k} \frac{i}{i!} \lambda^i \frac{\mu^{k-i}}{(k-i)!}$$

$$= e^{-(\lambda+\mu)} \lambda \sum_{i=1}^{k} \frac{1}{(i-1)!(k-i)!} \lambda^{i-1} \mu^{k-i}$$

$$= e^{-(\lambda+\mu)} \lambda \sum_{l=0}^{k-1} \frac{1}{l!(k-l-1)!} \lambda^{l} \mu^{k-l-1}$$

$l = i-1$

34

$$= e^{-(\lambda+\mu)} \frac{\lambda}{(k-1)!} \sum_{l=0}^{k-1} \underbrace{\frac{(k-1)!}{l!(k-l-1)!}}_{=\binom{k-1}{l}} \lambda^{l} \mu^{k-l-1}$$

$$= e^{-(\lambda+\mu)} \frac{\lambda}{(k-1)!} (\lambda+\mu)^{k-1}$$

same!

On the other hand,

$$\mathbb{E}\left(\frac{\lambda}{\lambda+\mu} Z ; Z=k\right) = \frac{\lambda k}{\lambda+\mu} \mathbb{P}(Z=k) = \frac{\lambda k}{\lambda+\mu} \frac{(\lambda+\mu)^k}{k!} e^{-(\lambda+\mu)}$$

→ What if the question was not to show that, but compute, not knowing the answer in advance?

General approach: find the CONDITIONAL DISTRIBUTION, and then compute the expectation under it.

It remains to define conditional distribution of Y given X.

① X is discrete.

Then we have already seen that $\mathbb{E}(Y|X) = \varphi(X)$, where $\varphi(x) = \dfrac{\mathbb{E}[Y \, \mathbb{1}(X=x)]}{\mathbb{P}(X=x)}$

If $\mathbb{P}(X=x) = 0$, $\varphi(x)$ is chosen arbitrarily.

$$= \sum_y y \boxed{\frac{\mathbb{P}(X=x, Y=y)}{\mathbb{P}(X=x)}}$$

define this guy as the conditional probability that $Y=y$ given $X=x$, and write

$$\mathbb{P}(Y=y \mid X=x) := \frac{\mathbb{P}(X=x, Y=y)}{\mathbb{P}(X=x)}$$

Compare with expression of expectation page 11.

② X has a density.

First, we note that characterization [CE.2] can be generalized to

[CE.2'] For any random variable Z which is $\sigma(X)$-measurable, $\mathbb{E}[\hat{Y} Z] = \mathbb{E}[YZ]$

Why? We know from [CE.2] that $\forall A \in \sigma(X)$, $\mathbb{E}[\hat{Y} \mathbb{1}_A] = \mathbb{E}[Y \mathbb{1}_A]$

instead $\mathbb{1}_A$, consider simple functions, then positive measurable functions, and finally any measurable function aka our random variable Z. We won't go into details.

So suppose that (X,Y) has joint density $f(x,y)$, and X has marginal distribution $f(x) = \int f(x,y)\,dy$.

Let $h: \mathbb{R} \to \mathbb{R}_+$ be a measurable function.

We compute $\mathbb{E}[h(Y)|X]$ the following way:

$$\mathbb{E}\left\{ \mathbb{E}[h(Y)|X]\, Z \right\} \overset{[CE.2']}{=} \mathbb{E}\left\{ h(Y)\, g(X) \right\}$$

Z is $\sigma(X)$ measurable
$\Rightarrow$ let's write it $g(X)$ for some measurable function g.

$$= \iint h(y)\, g(x)\, f(x,y)\, dx\, dy$$

$$= \int \left(\int h(y) f(x,y)\, dy \right) g(x)\, dx$$

$$= \int \left\{ \frac{\int h(y) f(x,y)\, dy}{f(x)} \right\} g(x)\, f(x)\, dx$$

define this quantity as $\varphi(x)$

$$= \int \varphi(x)\, g(x)\, f(x)\, dx$$

$$= \mathbb{E}[\varphi(X)\, g(X)]$$

$\Rightarrow \forall$ measurable $h: \mathbb{R} \to \mathbb{R}_+$, we just showed that

$$\mathbb{E}\left\{ \mathbb{E}[h(Y)|X]\, Z \right\} = \mathbb{E}\left\{ \varphi(X)\, Z \right\} \quad (\forall Z)$$

We conclude that $\mathbb{E}[h(Y)|X] = \varphi(X)$

So that $\mathbb{E}[h(Y)|X=x] = \varphi(x)$

$$= \int h(y) \boxed{\frac{f(x,y)}{f(x)}}\, dy$$

define this quantity as the conditional density of y given x, and write

$$f(y|x) = \frac{f(x,y)}{f(x)}.$$

In practice, to compute $E(Y|X)$, first compute $f(y|x)$, then compute $\varphi(x) = \int y\, f(y|x)\, dy$, and set $\mathbb{E}(Y|X) = \varphi(X)$

② Properties of CE.

[P1] Linearity. $\forall a, b \in \mathbb{R}$,

$$\mathbb{E}(aX + bY \mid Z) = a\,\mathbb{E}(X \mid Z) + b\,\mathbb{E}(Y \mid Z)$$

We need to show that the Right Hand Side (RHS) satisfies [CE.1] and [CE.2].

[CE.1] is trivially satisfies, since a function of Z.

Next,

$$\mathbb{E}[RHS\,;\, Z \in B] \underset{\text{linearity of } \mathbb{E}(.)}{=} a\,\mathbb{E}\big[\mathbb{E}(X \mid Z)\,;\, Z \in B\big] + b\,\mathbb{E}\big[\mathbb{E}(Y \mid Z)\,;\, Z \in B\big]$$

by definition of $\mathbb{E}(X|Z)$ and $\mathbb{E}(Y|Z)$

$$= a\,\mathbb{E}[X\,;\, Z \in B] + b\,\mathbb{E}[Y\,;\, Z \in B]$$

$$= \mathbb{E}[aX + bY\,;\, Z \in B]$$

so the RHS satisfies [CE.2]. ∎

[P2] Monotonicity

If $X \leq Y$ a.s then $\mathbb{E}(X \mid Z) \leq \mathbb{E}(Y \mid Z)$ a.s.

By contradiction, suppose the conclusion does not hold, so that

$$\mathbb{E}(Y - X \mid Z) \underset{[P1]}{=} \mathbb{E}(Y \mid Z) - \mathbb{E}(X \mid Z) < 0$$

with positive probability.

Put $\mathbb{E}(Y - X \mid Z) =: h(Z)$

[P1] + [P2] are analogous to the properties of $\mathbb{E}(.)$.

Next,

$$\{\omega \in \Omega : \mathbb{E}(Y - X \mid Z) < 0\} = \{\omega \in \Omega : h(Z) < 0\}$$

$$= \{\omega \in \Omega : Z \in B\}$$

where we defined $B := \{x \mid h(x) < 0\}$

By [CE.2],

$$\mathbb{E}[\underset{\substack{\text{strictly negative}\\ \text{on } B}}{h(Z)}\,;\, Z \in B] = \mathbb{E}[\underset{\substack{\text{positive}\\ \text{by assumption}}}{Y - X}\,;\, Z \in B]$$

set of positive probability
⇒ contradiction ∎

[P3] If $Y = g(Z)$ (measurable) then $\mathbb{E}(XY \mid Z) = Y\,\mathbb{E}(X \mid Z)$

The RHS is a function of Z so [CE.1] is satisfied.
Regarding [CE.2], we consider first the case of indicators $Y = \mathbb{1}(Z \in C)$ for $C \in \mathcal{B}(\mathbb{R})$.
Then

$$\to \mathbb{E}(XY\,;\, Z \in B) = \mathbb{E}\big(X\, \mathbb{1}(Z \in C)\, \mathbb{1}(Z \in B)\big)$$

$$\to \mathbb{E}(RHS\,;\, Z \in B) = \mathbb{E}\big(Y\,\mathbb{E}(X \mid Z)\,;\, Z \in B\big)$$

$$= \mathbb{E}\big(\mathbb{E}(X \mid Z)\,;\, Z \in B \times C\big)$$

[CE.2] for $\mathbb{E}(X|Z)$

$$= \mathbb{E}\big(X\,;\, Z \in B \times C\big),$$

which is the same as above.

Then move on to simple functions, their limits, etc. ∎

[P4] If X and Y are independent, then
$$\mathbb{E}(Y|X) = \mathbb{E}(Y).$$

$\mathbb{E}(Y)$ is a constant, which is a trivial function of X, so that [CE.1] is verified. Next,

$$\begin{aligned}
\mathbb{E}(Y\,;X\in B) &= \mathbb{E}(Y\,\mathbb{1}(X\in B)) \\
&\overset{\text{indpce}}{=} (\mathbb{E}\,Y)(\mathbb{E}\,\mathbb{1}(X\in B)) \\
&= \mathbb{E}\big((\mathbb{E}\,Y)\,\mathbb{1}(X\in B)\big) \\
&= \mathbb{E}(\mathbb{E}\,Y\,;X\in B) \\
&= \mathbb{E}(RHS\,;X\in B)
\end{aligned}$$
■

[P5] Double expectation law aka tower property.

$$\mathbb{E}\{\underbrace{\mathbb{E}(Y|X_1,X_2)}_{\text{less crude}}\mid X_1\} = \underbrace{\mathbb{E}(Y|X_1)}_{\text{more crude}}$$

Again, [CE.1] is immediate. For [CE.2],

$$\begin{aligned}
\mathbb{E}\{\mathbb{E}(Y|X_1,X_2)\,;X_1\in B\} &= \mathbb{E}\{\mathbb{E}(Y|X_1,X_2)\,;(X_1,X_2)\in B\times\mathbb{R}\} \\
&\overset{\substack{[CE.2] \\ \text{for } \mathbb{E}(Y|X_1,X_2)}}{=} \mathbb{E}\{Y\,;(X_1,X_2)\in B\times\mathbb{R}\} \\
&= \mathbb{E}\{Y\,;X_1\in B\} \\
&\overset{\substack{[CE.2] \\ \text{for } \mathbb{E}(Y|X_1)}}{=} \mathbb{E}\{\mathbb{E}(Y|X_1)\,;X_1\in B\} \\
&= \mathbb{E}\{RHS\,;X_1\in B\}
\end{aligned}$$
■

In particular, taking $X_1 = $ cst, $\mathbb{E}\{\mathbb{E}(Y|X)\} = \mathbb{E}\,Y$.

4 Property [P4] shows that if Y is independent of X, then $\mathbb{E}(Y|X) = \mathbb{E}(Y)$.
Show, however, that if for some X and Y such that $\mathbb{E}(Y|X) = \mathbb{E}(Y)$, this is not enough to conclude that X and Y are independent.

3 Geometric insights.

In the general definition of CE on page 31, the random variable Y is assumed to belong to $\mathcal{L}^1$. If in addition it is square integrable, then it is possible to take a different route to define the notion of conditional expectation. In fact, for $Y\in\mathcal{L}^2$, we show next that $\hat{Y} = \mathbb{E}(Y|X)$ is the best forecast of Y from X, under the metric $d^2(X,Y) = \mathbb{E}(Y-X)^2$.

Indeed, for a random variable $Z = \varphi(X)$ [your forecast should be a function of the predictor X]

$$\begin{aligned}
d^2(Y,Z) &= \mathbb{E}(Y-Z)^2 = \mathbb{E}\big((Y-\hat{Y})+(\hat{Y}-Z)\big)^2 \\
&= \mathbb{E}(Y-\hat{Y})^2 \\
&\quad + \mathbb{E}(\hat{Y}-Z)^2 \\
&\quad + 2\,\mathbb{E}(Y-\hat{Y})(\hat{Y}-Z)
\end{aligned}$$

We are looking for the $Z = \varphi(X)$ which minimizes this quantity

$$\mathbb{E}(Y-\hat{Y})(\hat{Y}-Z) \overset{[P5]}{=} \mathbb{E}\Big\{\mathbb{E}\big[(Y-\hat{Y})\underbrace{\boxed{(\hat{Y}-Z)}}_{\text{function of } X}\mid X\big]\Big\}$$

$\Rightarrow$ use [P3]

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

$$\mathbb{E}[\, Y - \hat{Y} \mid X \,] = \mathbb{E}(Y|X) - \mathbb{E}(\underbrace{\hat{Y}}_{\varphi(X)}|X) = \hat{Y} - \hat{Y} = 0$$

Thus, $\mathbb{E}(Y-Z)^2 = \underbrace{\mathbb{E}(Y-\hat{Y})^2}_{\text{independent of } Z} + \underbrace{\mathbb{E}(\hat{Y}-Z)^2}_{= 0 \text{ if and only if } Z = \hat{Y}.}$

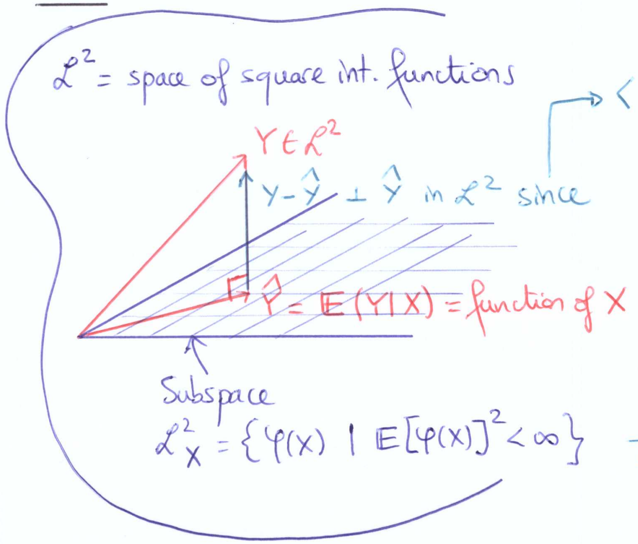
We just showed that

$$\boxed{\hat{Y} = \mathbb{E}(Y|X) = \underset{Z = \varphi(X)}{\operatorname{argmin}} \left\{ \mathbb{E}(Y-Z)^2 \right\}}$$

our best forecast is function of the covariate X

Mean square error $\equiv$ distance (square) between Y and its approximate value Z.

Picture:



$$\begin{aligned}
\langle \hat{Y}, Y - \hat{Y} \rangle &= \mathbb{E}[\hat{Y}(Y-\hat{Y})] \quad [P5] \\
&= \mathbb{E}\,\mathbb{E}[\hat{Y}(Y-\hat{Y}) \mid X] \quad [P3] \\
&= \mathbb{E}[\hat{Y}\, \underbrace{\mathbb{E}(Y-\hat{Y} \mid X)}_{= \mathbb{E}Y|X - \mathbb{E}\hat{Y}|X = 0}]
\end{aligned}$$

Thus $\langle \hat{Y}, Y - \hat{Y} \rangle = 0$

i.e. $\hat{Y} \perp Y - \hat{Y}$.