CONDITIONING ON Y = FIXING Y

First, for discrete Y:

Prop (Conditioning = fixing) If Y has a discrete distribution,

$$\mathbb{E}[X|Y] \overset{a.s.}{=} g(Y) \quad \text{where} \quad g(y) = \mathbb{E}[X|Y=y]$$

By Ex (6) p. 144,

$$\mathbb{E}[X|Y] = \underset{\substack{\parallel \\ g(y)}}{\mathbb{E}[X|F_y]} \text{ whenever the event } F_y := \{Y=y\} \text{ occurs}$$

- Prop can't hold for continuous distr since $\mathbb{P}\{Y=y\}=0 \ \forall y$. But:

Prop (Conditioning on Y fixes Y)

$$X \perp\!\!\!\perp Y \Rightarrow \mathbb{E}[f(X,Y)|Y] \overset{a.s.}{=} g(Y) \text{ where } g(y) = \mathbb{E} f(X,y).$$

i.e. "average w.r.to X while keeping Y fixed"

By def, suffices to check that

$$\mathbb{E}[f(X,Y)\underbrace{\mathbb{1}_{\{Y\in B\}}}_{=:\varphi(Y)}] = \mathbb{E}[g(Y)\mathbb{1}_{\{Y\in B\}}] \quad \forall B \in \mathcal{B}(\mathbb{R})$$

Realize (X,Y) as identity map on $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), \mathcal{L}_X \times \mathcal{L}_Y)$

$$= \int_{\mathbb{R}\times\mathbb{R}} f(x,y)\varphi(y)\, d\mathcal{L}_X(x)\, d\mathcal{L}_X(y) = \int_{\mathbb{R}} \underbrace{\left(\int_{\mathbb{R}} f(x,y)\, d\mathcal{L}_X(x)\right)}_{=g(y)} \varphi(y)\, d\mathcal{L}_Y(y)$$

distributions

Fubini

Alternatively, prove first for $f(X,Y) = \mathbb{1}_{\{X\in A, Y\in C\}}$,

then for simple functions, then take limit.

HW: Let Y be a r.v. $\quad H := \{X \in L^2 : X \perp\!\!\!\perp Y\}$

True/false?

(a) H is closed

(b) H is a linear subspace

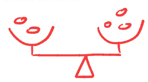
(c) $H \perp L_Y = \{$all L^2 r.v's of the form $h(Y)\}$

(d) $H =$ orth. complement of L_Y

Example 1.5.1 (Probability of a perfect cancelation)**.** Let $a_1, \ldots, a_n$ be real numbers, not all of which are zero. What is the probability that

$$\pm a_1 \pm \cdots \pm a_n = 0$$

Can be 1/2 ($a_1 = a_2$, rest=0)



scales

where the signs are chosen at random? Let us show that this probability is always bounded by 1/2. To state this problem rigorously, we can model the random signs as independent *Rademacher random variables* $X_1, \ldots, X_n$, i.e. random variables that take values -1 and 1 with probability $1/2$ each. We claim that

$$\mathbb{P}\{S_n = 0\} \leq \frac{1}{2} \quad \text{where} \quad S_n = \sum_{i=1}^{n} a_i X_i.$$

Proof We can assume without loss of generality that $a_n \neq 0$ by rearranging the terms if necessary. The proof is by exposing the last term $a_n X_n$ of the sum, while keeping all the previous terms fixed by conditioning. So let us condition on the random variables $X_1, \ldots, X_{n-1}$, or to be pedantic, on the random vector $(X_1, \ldots, X_{n-1})$. The conditioning fixes the values of $X_1, \ldots, X_{n-1}$, and therefore it fixes the value of $S_{n-1} = \sum_{i=1}^{n-1} a_i X_i$. All randomness is now left with X_n. Since $S_n = S_{n-1} + a_n X_n$, we have

$$\mathbb{P}\{S_n = 0 \,|\, X_1, \ldots, X_{n-1}\} = \mathbb{P}\Big\{X_n = -\frac{S_{n-1}}{a_n} \,\Big|\, X_1, \ldots, X_{n-1}\Big\} \leq \frac{1}{2}.$$

The inequality holds because X_n is independent of $X_1, \ldots, X_{n-1}$, the value $u = -S_{n-1}/a_n$ is fixed by conditioning, and the definition of the Rademacher distribution implies that $\mathbb{P}\{X_n = u\} \leq 1/2$ for any fixed value $u \in \mathbb{R}$.

We can finish the proof by applying the law of total expectation (1.16):

$$\mathbb{P}\{S_n = 0\} = \mathbb{E}\Big[\mathbb{P}\{S_n = 0 \,|\, X_1, \ldots, X_{n-1}\}\Big] \leq \mathbb{E}\,\frac{1}{2} = \frac{1}{2}. \quad \square$$

By the way, the result in Example 1.5.1 is sharp. If there are exactly two nonzero coefficients a_i and they are equal to each other, then we have $\mathbb{P}\{S_n = 0\} = 1/2$. (Check!)

Littlewood-Offord problem.

Ex: For Gaussian data, the best prediction is linear:

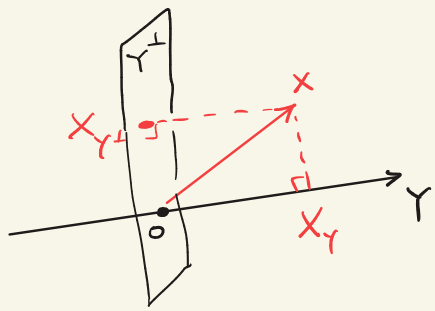
(Normal conditioning property)

Prop If X, Y are jointly normal, then $\mathbb{E}[X|Y] = aY + b$.

Proof WLOG $\mathbb{E}Y = 0$

Consider the L^2-orthogonal decomposition

$$X = X_Y + X_{Y^\perp}$$



where $X_Y := \langle X, \frac{Y}{\|Y\|_{L^2}} \rangle \frac{Y}{\|Y\|_{L_2}} = \underbrace{\frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}}_{=:a} \cdot Y$

$$\mathbb{E}[X|Y] = \mathbb{E}[\underbrace{X_Y}_{=aY}|Y] + \mathbb{E}[X_{Y^\perp}|Y]$$

$X_{Y^\perp} = X - aY$ and Y are jointly normal and orthogonal $\Rightarrow$ uncorrelated (since $\mathbb{E}Y=0$) $\Rightarrow$ independent.

$$= aY + \underbrace{\mathbb{E}[X_{Y^\perp}]}_{=\mathbb{E}(X-aY) = \mathbb{E}X}$$

$$= aY + \mathbb{E}X.$$

$\square$

Remark: the proof gives the explicit formula:

$$\mathbb{E}[X|Y] = \mathbb{E}X + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}(Y - \mathbb{E}Y)$$

HW: extend to: $\mathbb{E}[X|Y_1, \dots, Y_n] = a_1Y_1 + \cdots + a_nY_n + b$.

HW: Let X be normal, $h: \mathbb{R} \to \mathbb{R}$ be a mble function

Prove: $h(X)$ is normal $\Longleftrightarrow$ h is linear ?

Ex (Filtering)

Let $X \sim N(0,1)$ be an unknown signal, corrupted by unknown indep. noise $W \sim N(0,\sigma^2)$. You observe $Y = X + W$. Estimate X from Y, σ to minimize the mean-squared error.

- $\hat{X} := \operatorname{argmin}\{\mathbb{E}(\hat{X}-X)^2 : \hat{X} \text{ is } \sigma(Y)\text{-mble}\}$

$$\overset{\text{Cor p.151}}{=\!=} \mathbb{E}[X \mid Y]$$

$$\overset{\text{Remark}}{\underset{\text{prev. page}}{=\!=}} \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]} \cdot Y = \frac{\mathbb{E}X(X+W)}{\mathbb{E}(X+W)^2} \cdot Y = \boxed{\frac{Y}{1+\sigma^2}}$$

($\mathbb{E}X^2 = 1$, $\mathbb{E}XW = 0$ (indep), $X+W \sim N(0, 1+\sigma^2)$)

- $$\text{MSE} = \mathbb{E}(\hat{X}-X)^2 = \mathbb{E}\left(\frac{X+W}{1+\sigma^2} - X\right)^2 = \mathbb{E}\left(\frac{\overbrace{W-\sigma^2 X}^{N(0,\sigma^2+\sigma^4)}}{1+\sigma^2}\right)^2 = \frac{\sigma^2+\sigma^4}{(1+\sigma^2)^2} = \boxed{\frac{\sigma^2}{1+\sigma^2}}$$

- Interpretation:
  - Small noise $(\sigma \to 0) \Rightarrow \hat{X} \approx Y$, MSE $\to 0$ (trust observation)
  - Large noise $(\sigma \to \infty) \Rightarrow \hat{X} = 0$, MSE $\to 1$ (ignore observation)

- Normality is essential ($X \sim$ Rademacher, $W \sim \text{Unif}(-1,1)$ $\Rightarrow \hat{X} := \operatorname{sign}(X+W) = X$ is exact recovery)

HW extend to n noisy observations

$$Y_i \sim X + W_i$$

Show that $\hat{X} = \dfrac{\bar{Y}}{1+\bar{\sigma}^2}$ where $\bar{Y} = \frac{1}{n}\sum_1^n Y_i$, $\bar{\sigma}^2 = \frac{1}{n}\sigma_i^2$

$\underline{\underline{\text{Ex}}}$ (Predicting a term from a sum)

Let $X_1, \dots, X_n$ be iid r.v's, $S_n := X_1 + \dots + X_n$. Then

$$\mathbb{E}[X_i | S_n] = \frac{S_n}{n} \qquad \forall i \le n$$

⌜ $\mathbb{E}[X_i | S_n] =: f_i(S_n)$ don't depend on i by symmetry,

and $\sum_1^n f_i(S_n) = \mathbb{E}[S_n | S_n] = S_n \Rightarrow f_i = \frac{S_n}{n}$. ⌟

$\underline{\underline{\text{Ex}}}$ (Predicting a term from many sums)

$\underline{\underline{\text{Ex}}}$ $\mathcal{F}_n := \sigma(S_n, S_{n+1}, \dots) \Rightarrow \mathbb{E}[X_i | \mathcal{F}_n] = \frac{S_n}{n} \quad \forall i \le n$

⌜

$$\mathcal{F}_n \overset{\uparrow}{=} \sigma(S_n, X_{n+1}, X_{n+2}, \dots) = \sigma\Big(\underbrace{\sigma(S_n)}_{\mathcal{G}} \sqcup \underbrace{\sigma(X_{n+1}, \dots)}_{\mathcal{F}}\Big)$$

("$\subset$" from $S_k = S_n + X_{n+1} + \dots + X_k \quad \forall k \ge n$

"$\supset$" from $X_k = S_k - S_{k-1} \quad \forall k > n$)

$$\Rightarrow \mathbb{E}[X_i | \mathcal{F}_n] \overset{\text{p.149(c')}}{=\!=\!=} \mathbb{E}[X_i | \mathcal{G}] \overset{\text{Ex. above}}{=} \frac{S_n}{n}$$

⌟