ELE8088: Control & Estimation Theory **QUB, 2021**

Handout 11: Probability

Lecturer: Pantelis Sopasakis *Date:* _____

Topics: Probability spaces and modelling of random experiments ∘ Random variables ∘ Distribution ∘ Expectation ∘ Probability Density ∘ Normal and Uniform distributions ∘ Variance ∘ Multivariate random variables ∘ Multivariate normals ∘ Conditioning and independence.

Prerequisites: Handout X1 (linear algebra).

Last updated: June 4, 2022 at 20:53:22

11.1 Probability Spaces

11.1.1 Preliminaries: sets

Let A, B be two sets. We denote their union by $A \cup B$; it is $x \in A \cup B$ iff $x \in A$ or $x \in B$. We denote their intersection by $A \cap B$; it is $x \in A \cap B$ iff $x \in A$ and $x \in B$. We say that A is a subset of B (we denote $A \subseteq B$) if $x \in B$ whenever $x \in A$ (in other words, $x \in A$ implies $x \in B$). Lastly, suppose $A \subseteq \mathbb{R}^n$. The complement of A is denoted by A^c and it is the set of all elements of $\mathbb{R}^n$ that are not in A.

11.1.2 Probability spaces, events and probabilities

A *probability space* is a model that allows us to study random experiments such as the toss of a coin or the roll or a die. Let us start by considering a set Ω of all possible *outcomes* of a random experiments.

Next, we introduce the concept of an *event space*. An *event* is a set of outcomes, $A \subseteq \Omega$, to which we will later assign a probability value. The event space needs to satisfy certain axioms. For example, if A_1 and A_2 are two events, then $A_1 \cup A_2$ is also an event. Formally, the event space needs to be a σ-algebra. Let us give the definition[1].

Definition 11.1 (σ-algebra) *A σ-algebra, $\mathcal{F}$, on Ω is a collection of subsets of Ω such that*

1. *$\emptyset, \Omega \in \mathcal{F}$*
2. *If $A \in \mathcal{F}$, then its complement $A^c := \Omega \setminus A \in \mathcal{F}$*
3. *If $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$*

The collection $\mathcal{F} = \{\emptyset, \Omega\}$ is trivially a σ-algebra. The powerset of Ω, that is the set of all subsets of Ω, is a σ-algebra.

[1] The definition of a σ-algebra is not examinable

Next, we can define the *probability of an event* A, which is denoted by $\mathsf{P}[A]$. A *probability* P is a function that maps an event $A \in \mathcal{F}$ to a number in $[0,1]$ which satisfies three axioms which we are about to state. The probability of an event A can be thought of as a measure of the *likelihood* of the occurence of that event.

Definition 11.2 (Probability) *A probability* P *is a function that maps an event* $A \in \mathcal{F}$ *to a number in* $[0,1]$*, with*

1. $\mathsf{P}[\emptyset] = 0$ *and* $\mathsf{P}[\Omega] = 1$
2. $A \subseteq B \Rightarrow \mathsf{P}[A] \leq \mathsf{P}[B]$
3. *If* $A_1, A_2, \ldots$ *are mutually disjoint, then* $\mathsf{P}\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \mathsf{P}[A_i]$

Having given the definition of a σ-algebra and a probability, we can now give the definition of a probability space.

Definition 11.3 (Probability space) *A probability space consists of*

1. *a sample space, which is a nonempty set* Ω
2. *the set of events,* $\mathcal{F}$*, known as a* σ*-algebra of* Ω *(*$\mathcal{F} \subseteq 2^{\Omega}$*)*
3. *a probability* $\mathsf{P} : \mathcal{F} \to [0,1]$

In other words, a probability space is a triplet $(\Omega, \mathcal{F}, \mathsf{P})$*.*

Note that a pair $(\Omega, \mathcal{F})$ consisting of a sample space Ω and an event space (σ-algebra $\mathcal{F}$) is called a *measurable space.*

Example (Fair die). Consider a fair die. The sample space is

$$\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}\,. \tag{11.1}$$

Suppose that $\mathcal{F} = 2^\Omega$. For example,

$$\{\boxdot, \boxdot\} \in \mathcal{F},\ \emptyset \in \mathcal{F},\ \Omega \in \mathcal{F},\ \text{etc.} \tag{11.2}$$

Define P as the unique probability satisfying

$$\mathsf{P}[\{\omega\}] = \tfrac{1}{6}, \forall \omega \in \Omega. \tag{11.3}$$

Then,

$$\mathsf{P}[\{\boxdot, \boxdot\}] = \mathsf{P}[\{\boxdot\} \cup \{\boxdot\}] = \mathsf{P}[\{\boxdot\}] + \mathsf{P}[\{\boxdot\}] = \tfrac{1}{3}, \tag{11.4}$$

and in fact we can compute the probability of any subset of Ω. •

Exercise 1 (Probability of complement) ☕. Prove that $\mathsf{P}[A^c] = 1 - \mathsf{P}[A]$, for all $A \in \mathcal{F}$; we denote $A^c = \Omega \setminus A$. Using this property, determine the probability of rolling a fair die and *not* getting ⚅. ⋄

Exercise 2 (Probability of union and intersection) ☕☕. Prove that

1. $\mathsf{P}[A \cap B] \leq \min\{\mathsf{P}[A], \mathsf{P}[B]\}$, for all $A, B \in \mathcal{F}$
2. $\mathsf{P}[A \cup B] = \mathsf{P}[A] + \mathsf{P}[B] - \mathsf{P}[A \cap B]$ ⋄

Exercise 3 (De Morgan's Law #1) ☕☕. Show that

$$(A \cap B)^c = A^c \cup B^c. \tag{11.5}$$

Hint: Two sets are equal to one another if each is a subset of the other, i.e., $A = B$ means $A \subseteq B$ and $B \subseteq A$; show that $(A \cap B)^c \subseteq A^c \cup B^c$ and $A^c \cup B^c \subseteq (A \cap B)^c$. ⋄

Exercise 4 (De Morgan's Law #1) ☕☕. Show that

$$(A \cup B)^c = A^c \cap B^c. \diamond \tag{11.6}$$

Exercise 5 (Law of total probability) ☕☕. Suppose $A_1, A_2, \ldots, A_N \in \mathcal{F}$ are mutually disjoint events such that $\bigcup_{i=1}^N A_i = \Omega$. Then, for every $B \in \mathcal{F}$

$$\mathsf{P}[B] = \sum_{i=1}^{N} \mathsf{P}[B \cap A_i]. \diamond \tag{11.7}$$

Exercise 6 (Nested sequences of sets) ☕☕.

1. Let $(A_i)_{i\in\mathbb{N}}$ be an nondecreasing sequence of events, i.e., $A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$. Show that

$$\mathsf{P}\left[\bigcup_{i=1}^{\infty} A_i\right] = \lim_{i\to\infty} \mathsf{P}[A_i] \tag{11.8}$$

   Hint: use the third axiom in the definition of probability and try to write the above union as a union of disjoint sets.

2. Let $(A_i)_{i\in\mathbb{N}}$ be an nonincreasing sequence of events, i.e., $A_1 \supseteq A_2 \supseteq A_3 \supseteq \ldots$. Show that

$$\mathsf{P}\left[\bigcap_{i=1}^{\infty} A_i\right] = \lim_{i\to\infty} \mathsf{P}[A_i] \tag{11.9}$$

   Hint: use the result of the previous question. $\diamond$

11.1.3 Random variables

Suppose you place a bet — if you roll a ⚃ you win £10, if you roll ⚀ you lose £100 — otherwise, you don't win anything. Your winnings are:

$$X(\omega) = \begin{cases} \text{£}10, & \text{if } \omega = ⚃ \\ -\text{£}100, & \text{if } \omega = ⚀ \\ \text{£}0, & \text{otherwise} \end{cases} \tag{11.10}$$

Function $X : \Omega \to \mathbb{R}$ is a (real-valued) *random variable*. Note that ω is decided by nature or, in general, by a mechanism that is unknown to us.

Definition 11.4 (Random variable) *Given a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ and a measurable space $(E, \mathcal{E})$ (sample space + σ-algebra), a random variable X is a function $X : \Omega \to E$ such that $X^{-1}(B) \in \mathcal{F}$, for all $B \in \mathcal{E}$.*

In other words, a random variable is a function that inverts sets of $\mathcal{E}$ to sets of $\mathcal{F}$; the reason for this requirement will become evident in a while.

Typically, E can be (i) a discrete space, (ii) the set of real numbers, $\mathbb{R}$, or (iii) the set of real vectors. For example, in the case of the variable in Equation (11.10) we have $E = \{-100, 0, 10\}$.

Suppose you place a bet and your winnings are

$$X(\omega) = \begin{cases} £1, & \text{if } \omega \in \{\boxdot, \boxdot, \boxdot\} \\ £2, & \text{if } \omega \in \{\boxdot, \boxdot, \boxdot\} \end{cases} \tag{11.11}$$

What is the probability of your winnings being £2? It is

$$\begin{aligned} \mathsf{P}[X = £2] &= \mathsf{P}[\{\omega \in \Omega : X(\omega) = £2\}] \\ &= \mathsf{P}[\{\boxdot, \boxdot, \boxdot\}] = 1/2. \end{aligned}$$

In general, define the *probability distribution* of X as the function $F_X : \mathcal{E} \to [0, 1]$ given by

$$F_X(B) = \mathsf{P}[X \in B] = \mathsf{P}[\{\omega \in \Omega : X(\omega) \in B\}] = \mathsf{P}[X^{-1}(B)]. \tag{11.12}$$

Note that since X is a random variable, $X^{-1}(B) \in \mathcal{F}$, so $\mathsf{P}[X^{-1}(B)]$ is well defined.

11.1.4 Expectation

Suppose you place a bet: if you roll a ⚅ you win £10, if you roll ⚀ you lose £100 — otherwise, you win £5 . Your winnings are:

$$X(\omega) = \begin{cases} £10, & \text{if } \omega = \boxdot \\ -£100, & \text{if } \omega = \boxdot \\ £5, & \text{otherwise} \end{cases} \tag{11.13}$$

The possible values of X are $X(\Omega) = \{10, -100, 5\}$. Your *expected winnings* are defined as

$$\begin{aligned} \mathbb{E}[X] &= \sum_{\omega \in \Omega} X(\omega)\mathsf{P}[\{\omega\}] \\ &= \underbrace{(-100)\tfrac{1}{6}}_{\boxdot} + \underbrace{5\tfrac{1}{6}}_{\boxdot} + \underbrace{5\tfrac{1}{6}}_{\boxdot} + \underbrace{5\tfrac{1}{6}}_{\boxdot} + \underbrace{5\tfrac{1}{6}}_{\boxdot} + \underbrace{10\tfrac{1}{6}}_{\boxdot} \approx -£11.67, \end{aligned} \tag{11.14}$$

and $\mathbb{E}[X]$ is the *expectation* of X. Note that this definition can only be applied when Ω is *discrete* (finite or countably infinite).

Let us see what to expect when the space is countably infinite: Suppose $\Omega = \mathbb{N}$ and let $X : \Omega \to \mathbb{R}$ be a random variable. Then, analogously

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega)\mathsf{P}[\{\omega\}]. \tag{11.15}$$

Example (Expectation of Poisson distribution). Suppose that for $k \in \mathbb{N}$

$$\mathsf{P}[\{k\}] = \frac{\lambda^k e^{-\lambda}}{k!}, \tag{11.16}$$

where $\lambda > 0$ is a constant. Suppose $X(k) = k$; then

$$\mathbb{E}[X] = \sum_{k \in \mathbb{N}} X(k)\mathsf{P}[\{k\}] = \sum_{k \in \mathbb{N}} k\frac{\lambda^k e^{-\lambda}}{k!} = \ldots = \lambda. \bullet \tag{11.17}$$

Let us give the definition of the expectation of a discrete random variable.

Definition 11.5 (Expectation of discrete random variable) *Let $X : (\Omega, \mathcal{F}, \mathsf{P}) \to E$ be a discrete random variable and $E \subseteq \mathbb{R}^n$. The expectation of X is defined as*

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega)\mathsf{P}[\{\omega\}]. \tag{11.18}$$

Exercise 7 (☕). We flip a coin twice. If we get the same outcome (e.g., two heads or two tails), we earn £1. If we get opposite outcomes (i.e., a head and a tail), we lose 50p. Model this experiment by introducing a probability space and a random variable. Determine the expected earnings. ⋄

11.2 Density or Real-Valued Random Variables

We start by giving the definition of two important constructs that allow us to study real-valued random variables: the cumulative distribution function (cdf) and the probability density function (pdf). Later we will generalise these definitions to vector-valued random variables.

Definition 11.6 (CDF and PDF of real-valued random variables) *Let X be a real-valued random variable. The cumulative distribution function (cdf) of X is a function $F_X : \mathbb{R} \to [0, 1]$ defined as*

$$F_X(x) = \mathsf{P}[X \leq x]. \tag{11.19}$$

A function $p_X : \mathbb{R} \to [0, \infty)$ is called the probability density function (pdf) of X if

$$\mathsf{P}[a \leq X \leq b] = \int_a^b p_X(x)\mathrm{d}x. \tag{11.20}$$

If X has a pdf[2], it is called a *continuous* random variable. Using the pdf of X we can determine the probability that $X \in A$, where $A \subseteq \mathbb{R}$, as follows

$$\mathsf{P}[X \in A] = \int_A p_X(x)\mathrm{d}x. \tag{11.21}$$

By definition of the cdf — see Equation (11.19) — is given by

$$F_X(x) = \int_{-\infty}^{x} p_X(\xi)\mathrm{d}\xi. \tag{11.22}$$

By the fundamental theorem of calculus (FTC), if p_X is continuous at x, then

$$p_X(x) = F'_X(x). \tag{11.23}$$

Note that

$$\mathsf{P}[a \leq X \leq b] = F_X(b) - F_X(a). \tag{11.24}$$

A cdf, F_X, is well defined if

[2] All real-valued random variables have a cdf, but not all have a pdf

- F_X is right-continuous
- F_X is non-decreasing (if $x_1 \leq x_2$, then $F_X(x_1) \leq F_X(x_2)$)
- $\lim_{x\to-\infty} F_X(x) = 0$
- $\lim_{x\to\infty} F_X(x) = 1$

A pdf, p_X, of a real-valued random variable, X, is well defined if

- $p_X(x) \geq 0$ for all $x \in \mathbb{R}$
- $\int_{-\infty}^{\infty} p_X(\xi)\mathrm{d}\xi = 1$

Let us give a couple of examples of continuous random variables.

Example (Normal distribution). If a real-valued random variable X has the following pdf

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \tag{11.25}$$

with $\mu \in \mathbb{R}$, $\sigma > 0$, we say that X follows the *normal distribution* $\mathcal{N}(\mu, \sigma^2)$ and we denote this by $X \sim \mathcal{N}(\mu, \sigma^2)$. •

Example (Uniform distribution). If X has the following pdf

$$p_X(x) = \begin{cases} \frac{1}{b-a}, & \text{for } a \leq x \leq b \\ 0, & \text{otherwise} \end{cases} \tag{11.26}$$

where $a < b$, we say that X follows the *uniform distribution* $U(a, b)$. •

11.2.1 Notable properties of pdfs

Some notable properties of a probability density function:

1. By definition

$$\mathsf{P}[a \leq X \leq b] = \int_a^b p_X(x)\mathrm{d}x \tag{11.27}$$

2. By taking $a = b$[3]

$$\mathsf{P}[X = a] = 0 \tag{11.28}$$

3. Since $\mathsf{P}[-\infty < X < \infty] = 1$, it is

$$\int_{-\infty}^{\infty} p_X(x)\mathrm{d}x = 1. \tag{11.29}$$

4. We have defined $F_X(x) = \mathsf{P}[X \leq x]$, so

$$F_X(x) = \int_{-\infty}^{x} p_X(x)\mathrm{d}x \tag{11.30}$$

5. If p_X is continuous at x, then

$$F'_X(x) = p_X(x) \tag{11.31}$$

Exercise 8 (PDF of Uniform Distribution) ☕. Show that the pdf of $U([a, b])$, with $a < b$, is well defined in the sense that (i) $p_X(x) \geq 0$, (ii) $\int_{-\infty}^{\infty} p_X(x)\mathrm{d}x = 1$. ⋄

Exercise 9 (PDF of Uniform Distribution) ☕☕. Similarly, show that the exponential distribution with parameter $\lambda > 0$, which is given by

$$p_X(x) = \begin{cases} 0, & \text{for } x < 0 \\ \lambda e^{-\lambda x}, & \text{for } x \geq 0 \end{cases} \tag{11.32}$$

is well defined. ⋄

Exercise 10 (PDF of Pareto Distribution) ☕☕. The Pareto distribution with scale parameter $\alpha > 0$ and shape parameter $x_0 > 0$, has the pdf

$$p_X(x) = \begin{cases} 0, & \text{for } x < x_0 \\ \lambda \frac{\alpha x_0^{\alpha}}{x^{\alpha+1}}, & \text{for } x \geq x_0 \end{cases} \tag{11.33}$$

Show that this pdf is well defined. ⋄

Exercise 11 (PDF of Normal Distribution) ☕☕☕. The pdf of the normal distribution, $\mathcal{N}(\mu, \sigma^2)$, with $\sigma^2 > 0$ is given in Equation (11.25). Show that this pdf is well defined. ⋄

[3]Note that this propery holds for *continuous* random variables, but as we saw earlier this is not necessarily the case for discrete random variables.

Exercise 12 (Weibull distribution) ☕☕. The Weibull distribution is widely used in reliability engineering to study the failure rate of equipment. The pdf of a Weibull distribution with scale parameter $\lambda > 0$ and shape parameter $k > 0$ is given by

$$p_X(x) = \begin{cases} 0, & \text{for } x < 0 \\ \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \exp\left[-\left(\frac{x}{\lambda}\right)^k\right], & \text{for } x \geq 0 \end{cases} \tag{11.34}$$

Show that the cdf is

$$F_X(x) = \begin{cases} 0, & \text{for } x < 0 \\ 1 - \exp\left[-\left(\frac{x}{\lambda}\right)^k\right], & \text{for } x \geq 0 \end{cases} \quad \diamond \tag{11.35}$$

The expectation of a continuous real-valued random variable X can be determined using the pdf of the random variable. Let us give the following definition[4]

Definition 11.7 (Expectation of continuous random variable) *Let X be a continuous real-valued random variable with pdf p_X. The expectation of X is given by*

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x p_X(x) \mathrm{d}x, \tag{11.36}$$

provided that the integral converges.

Example (Uniform distribution). If $X \sim U(a, b)$, then

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x p_X(x) \mathrm{d}x = \int_a^b x \tfrac{1}{b-a} \mathrm{d}x = \tfrac{a+b}{2}. \bullet \tag{11.37}$$

Example (Normal distribution). If $X \sim \mathcal{N}(\mu, \sigma^2)$, then[5]

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \tfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}x = \ldots = \mu. \bullet$$

[4]Strictly speaking, this is not a definition. Note that we have defined the expectation of *discrete* and *continuous* random variables, but not the expectation of a general random variable. This is a deliberate choice for two reasons: (i) we will only be working with discrete and continuous random variables, (ii) the definition in the more general case requires a lengthy introduction.

[5]Hint: $\int_{-\infty}^{\infty} e^{-x^2} \mathrm{d}x = \sqrt{\pi}$.

Exercise 13 (Exponential distribution) ☕☕. Suppose X follows the exponential distribution with parameter $\lambda > 0$, i.e., $p_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ and $p_X(x) = 0$ for $x < 0$. Show that $\mathbb{E}[X] = 1/\lambda$. $\diamond$

Some notable properties of the expectation are:

1. If X is a discerete random variable on a space $\Omega = \{1, \ldots, n\}$, by definition

$$\mathbb{E}[X] = \sum_{i=1}^{n} p_i X_i, \tag{11.38}$$

   where $p_i = \mathsf{P}[\{i\}]$ and $X_i = X(\omega_i)$

2. If X is continuous

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x p_X(x) \mathrm{d}x. \tag{11.39}$$

3. $\mathbb{E}$ is linear: for random variables X, Y for which $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ exist and are finite, and $a, b \in \mathbb{R}$

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]. \tag{11.40}$$

4. If $\mathsf{P}[X < 0] = 0$, then $\mathbb{E}[X] \geq 0$. If $\mathsf{P}[X < 0] = 0$ we say that $X \geq 0$ *almost surely.*

5. (Law of the Unconscious Statistician — for short, *LotUS*[6]) If X is a continuous RV with pdf p_X and $g : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) p_X(x) \mathrm{d}x. \tag{11.41}$$

11.2.2 Variance

The variance of a real-valued (discrete or continuous) random variance is a measure of its dispersion/spread. Let us start by stating the definition.

[6] This result is known as the law of the unconscious statistician because often people (statisticians, allegedly) tend to treat it as an axiomatically correct without realising that it is actually a theorem

Definition 11.8 (Variance) *The variance of a real-valued (discrete or continuous) random variable is defined as*

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]. \tag{11.42}$$

Equivalently,

$$\begin{aligned}\mathrm{Var}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2. \end{aligned} \tag{11.43}$$

If X is a discrete random variable on a space $\Omega = \{1, \ldots, n\}$, with expectation $\mu = \mathbb{E}[X]$, then

$$\mathrm{Var}[X] = \sum_{i=1}^{n} p_i (X_i - \mu)^2 = \sum_{i=1}^{n} p_i X_i^2 - \mu^2. \tag{11.44}$$

Example (Variance of Discrete Random Variable). Let X be the outcome of a fair die roll, i.e., $\Omega = \{1, \ldots, 6\}$, $X(\omega) = \omega$ and denote $X_i = i$, $p_i = \frac{1}{6}$. Then,

$$\mathbb{E}[X] = \sum_{i=1}^{6} p_i X_i = \sum_{i=1}^{6} i \cdot \tfrac{1}{6} = 3.5, \tag{11.45}$$

and

$$\mathrm{Var}[X] = \sum_{i=1}^{n} p_i X_i^2 - \mu^2 = 2.9167. \bullet \tag{11.46}$$

Example (Variance of Uniform Distribution). Let $X \sim U(a, b)$. We know that $\mathbb{E}[X] = \frac{b-a}{2}$; then,

$$\begin{aligned}\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] &\overset{\text{LotUS}}{=} \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 \, p_X(x) \mathrm{d}x \\ &= \int_a^b \left(x - \tfrac{a+b}{2}\right)^2 \tfrac{1}{b-a} \mathrm{d}x \\ &= \tfrac{1}{b-a} \int_a^b \left(x - \tfrac{a+b}{2}\right)^2 \mathrm{d}x = \ldots = \tfrac{1}{12}(b-a)^2, \end{aligned} \tag{11.47}$$

where in the second equation we used LotUS. •

Example (Variance of Normal Distribution). Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then,

$$\begin{aligned}\mathrm{Var}[X] = \mathbb{E}[(X-\mu)^2] &= \int_{-\infty}^{\infty} (x-\mu)^2 p_X(x) \mathrm{d}x \\ &= \tfrac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x = \ldots = \sigma^2. \bullet \end{aligned} \tag{11.48}$$

Exercise 14 (Variance of Normal Distribution) ☕☕☕. Work out the integral in Equation (11.48). Hints: (i) Write p_X as follows

$$p_X(x) = \tfrac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2}, \tag{11.49}$$

(ii) Use the change of variables $u = \frac{x-\mu}{\sqrt{2}\sigma}$, (iii) Apply integration by parts, (iv) Use the fact that $\int_{-\infty}^{\infty} e^{-u^2} \mathrm{d}u = \sqrt{\pi}$. ⋄

Some properties of the variance:

1. The variance is nonnegative
2. For every $c \in \mathbb{R}$, $\mathrm{Var}[X + c] = \mathrm{Var}[X]$
3. For every $a \in \mathbb{R}$, $\mathrm{Var}[aX] = a^2 \mathrm{Var}[X]$

Exercise 15 (Proof) ☕. Prove the above properties. ⋄

Exercise 16 (Variance of scaled random variable) ☕. Let X be a real-valued random variable and $a \in \mathbb{R}$. Then $\mathrm{Var}[aX] = a^2 \mathrm{Var}[X]$. ⋄

11.3 Multivariate Random Variables

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space. A *multivariate random variables* is a function $X : (\Omega, \mathcal{F}, \mathsf{P}) \to \mathbb{R}^n$, i.e.,

$$X(\omega) = \begin{bmatrix} X_1(\omega) & \ldots & X_n(\omega) \end{bmatrix}^\mathsf{T}. \tag{11.50}$$

The *expectation* of X is defined as

$$\mathbb{E}[X] = \begin{bmatrix} \mathbb{E}[X_1] & \ldots & \mathbb{E}[X_n] \end{bmatrix}^\mathsf{T}. \tag{11.51}$$

We will now generalise the definitions of cdf and pdf functions to multivariate random variables.

Definition 11.9 (cdf and pdf of multivariate random variable) *The cumulative distribution function (cdf) of a multivariate random variable, $X : (\Omega, \mathcal{F}, \mathsf{P}) \to \mathbb{R}^n$, is a function $F_X : \mathbb{R}^n \to [0,1]$ defined by*

$$F_X(x_1, x_2, \ldots, x_n) = \mathsf{P}[X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n]. \tag{11.52}$$

We say that a function $p_X : \mathbb{R}^n \to \mathbb{R}$ is the probability density function of X if

$$\mathsf{P}[X \in A] = \int_A p_X(x)\mathrm{d}x. \tag{11.53}$$

In Section 11.2.2 we defined the variance of real-valued random variables; the variance is a nonnegative scalar that quantifies that spread of the random variable. When dealing with vector-valued random variables, the counterpart of the variance is the *variance covariance matrix*, which is a symmetric positive semidefinite matrix. Let us give the definition.

Definition 11.10 (Variance-covariance martrix) *The variance-covariance matrix of an $\mathbb{R}^n$-valued random variable X is defined as*

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\intercal]. \tag{11.54}$$

Note that in Equation (11.54), $X - \mathbb{E}[X]$ is an $\mathbb{R}^n$-valued random variable, so $(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\intercal$ is an $n \times n$ matrix and $\mathrm{Var}[X]$ is defined as the expectation of a matrix. The expectation of a matrix is defined as the expectation of its elements; this is in line with Equation (11.51).

Exercise 17 (Variance-covariance martrix) ☕. Show that the variance-covariance matrix of an $\mathbb{R}^n$-valued random variable X is given by

$$\mathrm{Var}[X] = \mathbb{E}[XX^\intercal] - \mathbb{E}[X]\mathbb{E}[X]^\intercal. \diamond \tag{11.55}$$

Definition 11.11 (Cross-covariance martrix) *Given two n-dimensional random variables X, Y, their cross-covariance matrix (aka covariance matrix) is*

$$\mathrm{Cov}[X,Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^\intercal]. \tag{11.56}$$

Exercise 18 (Cross-covariance martrix) ☕. Show that the cross-covariance matrix of two $\mathbb{R}^n$-valued random variables X and Y is given by

$$\mathrm{Cov}[X,Y] = \mathbb{E}[XY^\intercal] - \mathbb{E}[X]\mathbb{E}[Y]^\intercal. \diamond \tag{11.57}$$

Example (Multivariate case: probability, expectation and marginal pdf). Let X, Y be two real-valued continuous random variables and $Z = (X, Y)$. Let

$$A = \left\{ \left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right] \in \mathbb{R}^2 : a \le x \le b, c \le y \le d \right\}. \tag{11.58}$$

Then

$$\mathsf{P}[Z \in A] = \int_A p_Z(z)\mathrm{d}z = \int_c^d \int_a^b p_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y. \tag{11.59}$$

The expectation of Z is

$$\mathbb{E}[Z] = \int_{\mathbb{R}^2} z p_Z(z)\mathrm{d}z = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right] p_Z(x,y)\mathrm{d}x\mathrm{d}y. \tag{11.60}$$

The expectation of X is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x p_X(x)\mathrm{d}x, \tag{11.61}$$

where $p_X(x)$ is the *marginal pdf* of X, which is

$$p_X(x) = \int_{-\infty}^{\infty} p_Z(x,y)\mathrm{d}y. \bullet \tag{11.62}$$

Example (Marginal pdf). Consider an $\mathbb{R}^2$-valued continuous random variable $X = (X_1, X_2)$ with

$$p_X(x) = \begin{cases} 3x_1 + 1, & \text{for } x_1, x_2 \ge 0, \text{ and } x_1 + x_2 \le 1 \\ 0, & \text{otherwise} \end{cases} \tag{11.63}$$

The reader can verify that p_X is well defined. We will determine the marginal pdf $p_{X_1}(x_1)$. Using the definition of the marginal pdf in Equation (11.62) we have

$$p_{X_1}(x_1) = \int_{-\infty}^{+\infty} p_X(x_1, x_2)\mathrm{d}x_2 = \int_0^{1-x_1} (3x_1 + 1)\mathrm{d}x_2 = (3x_1 + 1)(1 - x_1), \tag{11.64}$$

defined for $x_1 \in [0, 1]$, and $p(x_1) = 0$ for $x \notin [0, 1]$. •

Expectation of multivariate random variables: $\mathbb{E}$ is linear: for random variables X, Y for which $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ exist and are finite, and $a, b \in \mathbb{R}$

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]. \tag{11.65}$$

Exercise 19 (Expectation of linear transformation) ☕. Use the linearity property of $\mathbb{E}$ (Equation (11.65)) to show that for a matrix $A \in \mathbb{R}^{m\times n}$,

$$\mathbb{E}[AX] = A\mathbb{E}[X]. \diamond \tag{11.66}$$

Exercise 20 (Variance of linear transformation) ☕. Show that

$$\mathrm{Var}[AX] = A\,\mathrm{Var}[X]A^\intercal. \diamond \tag{11.67}$$

Exercise 21 (Optimality of expectation) ☕☕☕.: (i) Given an n-dimensional RV X, define the function $f : \mathbb{R}^n \to \mathbb{R}$

$$f(z) = \mathbb{E}[\|X - z\|^2]. \tag{11.68}$$

Then $f(z) \geq f(\mathbb{E}[X])$ for all $z \in \mathbb{R}^n$. In other words, $\mathbb{E}[X] \in \arg\min_z f(z)$.

(ii) Define the function $F : \mathbb{R}^n \to \mathbb{S}^n_{++}$

$$F(z) = \mathbb{E}[(X - z)(X - z)^\intercal]. \tag{11.69}$$

Then $F(z) \succcurlyeq F(\mathbb{E}[X])$ for all $z \in \mathbb{R}^n$. We can say that, in a way, $\mathbb{E}[X]$ "minimises" F.

(iii) Use (ii) to prove (i); hint: use the property $\operatorname{trace}\mathbb{E}[X] = \mathbb{E}[\operatorname{trace}(X)]$. $\diamond$

11.3.1 Multivariate normal distribution

Suppose that $Z_1, \ldots, Z_m \sim \mathcal{N}(0, 1)$ and $A \in \mathbb{R}^{n\times m}$, $\mu \in \mathbb{R}^n$. Let

$$X = AZ + \mu. \tag{11.70}$$

We say that X follows the *multivariate normal distribution* $\mathcal{N}(\mu, \Sigma)$, where $\Sigma = AA^\intercal$.

The expectation of X is μ and its variance-covariance matrix is $\mathrm{Var}[X] = \Sigma$.

If Σ is positive definite, then the pdf of $\mathcal{N}(\mu, \Sigma)$ is

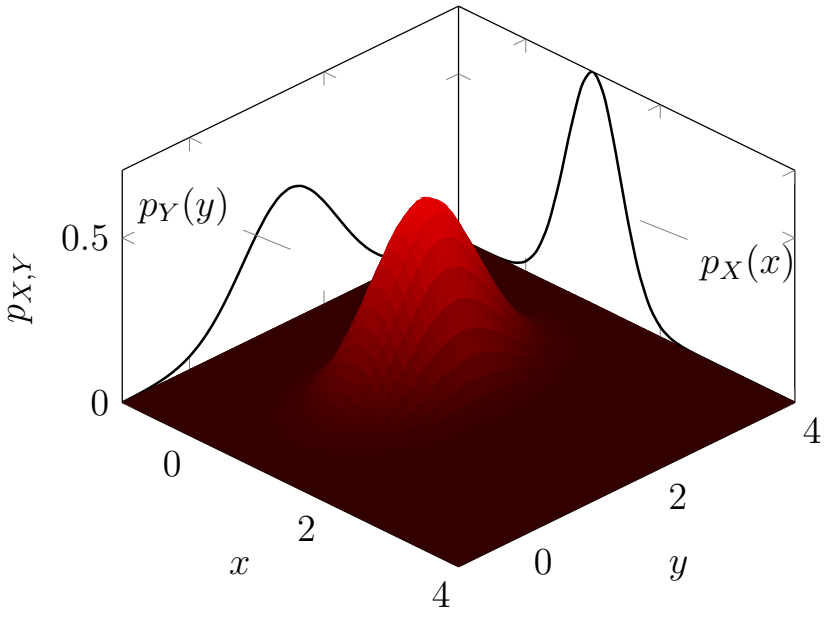
$$p_X(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{n/2}} e^{-\frac{1}{2}(x-\mu)^\intercal \Sigma^{-1}(x-\mu)}. \tag{11.71}$$

Marginals: Suppose $Z = (X, Y) \sim \mathcal{N}(\mu, \Sigma)$ with $\mu = (\mu_X, \mu_Y)$ and

$$\Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}, \tag{11.72}$$

then

$$X \sim \mathcal{N}(\mu_X, \Sigma_{XX}). \tag{11.73}$$



Affine transformations: We can easily show that for any matrix $A \in \mathbb{R}^{m\times n}$, vector $b \in \mathbb{R}^n$, and n-dimensional random variable X, we have

$$\mathbb{E}[AX + b] = A\mathbb{E}[X] + b. \tag{11.74}$$

Therefore, if $X \sim \mathcal{N}(\mu, \Sigma)$,

$$\mathbb{E}[AX + b] = A\mu + b, \tag{11.75}$$

and

$$\mathrm{Var}[AX + b] = \mathbb{E}[(AX - A\mu)(AX - A\mu)^\intercal] = A\,\mathrm{Var}[X]A^\intercal = A\Sigma A^\intercal. \tag{11.76}$$

Moreover, we can show that

$$AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\intercal). \tag{11.77}$$

11.4 Understanding conditioning on discrete spaces

11.4.1 Conditional probability

Here are some examples of the concept of *conditional probability* or *probability conditioned by an event*

- What is the probability that a person develops prostate cancer *given that* they are male *and* smoke? This is denoted by $\mathsf{P}[\text{Cancer} \mid \text{Male}, \text{Smoker}]$
- What is the probability that a student scores $> 90\%$ in the exam *given that* they study for 1 hr every day? This is denoted by $\mathsf{P}[\text{Score} > 90\% \mid \text{Study 1 hour every day}]$
- What is the probability that a person has COVID-19 *given that* they have fever? This is denoted by $\mathsf{P}[\text{COVID19} \mid \text{Fever}]$
- What is the probability that a person has fever *given that* they have COVID-19? This is denoted by $\mathsf{P}[\text{Fever} \mid \text{COVID19}]$

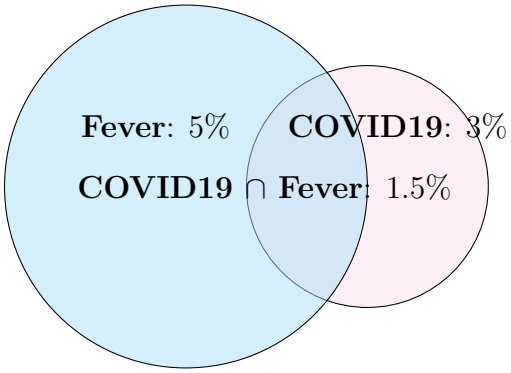
This leads to the introduction of the concept of the *conditional probability.* Let us give the definition.

Definition 11.12 (Conditional probability) *Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space and $A, B \in \mathcal{F}$ and $\mathsf{P}[B] > 0$. We define*

$$\mathsf{P}[A \mid B] = \frac{\mathsf{P}[A \cap B]}{\mathsf{P}[B]}. \tag{11.78}$$

This is called the conditional probability of A given B.

If $\mathsf{P}[A \mid B] = \mathsf{P}[A]$, we say that A and B are *independent.* If A and B are independent, then $\mathsf{P}[A \cap B] = \mathsf{P}[A]\mathsf{P}[B]$. We will revisit the concept of independence in Section 11.5.3.



$$\mathsf{P}[\text{Fever} \mid \text{COVID19}] = \frac{\mathsf{P}[\text{Fever} \cap \text{COVID19}]}{\mathsf{P}[\text{COVID19}]} = \frac{0.015}{0.03} = 50\%$$

$$\mathsf{P}[\text{COVID19} \mid \text{Fever}] = \frac{\mathsf{P}[\text{Fever} \cap \text{COVID19}]}{\mathsf{P}[\text{Fever}]} = \frac{0.015}{0.05} = 30\%$$

Example (Conditioning of die rolls). We roll a fair die and we play a game where our winnings are

$$X(\omega) = \begin{cases} £0, & \text{if } \omega \in \{\boxdot, \boxdot, \boxdot\} \\ £1, & \text{if } \omega \in \{\boxdot, \boxdot, \boxdot\} \end{cases} \tag{11.79}$$

If we know that we have won £1, then the probability they rolled $\boxdot$ is zero, $\mathsf{P}[\{\boxdot, \boxdot, \boxdot\} \mid X = £1] = 0$, and $\mathsf{P}[\{\boxdot, \boxdot, \boxdot\} \mid X = £1] = 1$. What is the probability $\mathsf{P}[A \mid X = £1]$ for some event A? •

This gives rise to the probability of an event conditional on the outcome of a *discrete* random variable. Let us give the definition.

Definition 11.13 (Conditioning by discrete random variable) *x Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a discrete probability space and X is a (discrete) random variable. For $A \in \mathcal{F}$ we define*

$$\mathsf{P}[A \mid X = x] = \mathsf{P}[A \mid B(x)], \tag{11.80}$$

where $B(x) = \{\omega \in \Omega : X(\omega) = x\}$, provided that $\mathsf{P}[B(x)] > 0$.

We will discuss the case of conditioning by a continuous random variable in Section 11.5. In this section we will be dealing with discrete random variables only. Keep in mind that Definition 11.13 does not hold for continuous random variables; the reason is that if X is continuous, $\mathsf{P}[X = x] = \mathsf{P}[B(x)] = 0$. Let us give an example.

Example (Conditional probability of discrete random variable). Consider again the previous example where

$$X(\omega) = \begin{cases} £0, & \text{if } \omega \in \{\boxdot, \boxdot, \boxdot\} \\ £1, & \text{if } \omega \in \{\boxdot, \boxdot, \boxdot\} \end{cases} \tag{11.81}$$

Then,

$$\begin{aligned} \mathsf{P}[\{\boxdot, \boxdot\} \mid X = £1] &= \mathsf{P}[\{\boxdot, \boxdot\} \mid \{\omega \in \Omega : X(\omega) = 1\}] \\ &= \mathsf{P}[\{\boxdot, \boxdot\} \mid \{\boxdot, \boxdot, \boxdot\}] \\ &= \frac{\mathsf{P}[\{\boxdot, \boxdot\} \cap \{\boxdot, \boxdot, \boxdot\}]}{\mathsf{P}[\{\boxdot, \boxdot, \boxdot\}]} \\ &= \frac{\mathsf{P}[\{\boxdot\}]}{\mathsf{P}[\{\boxdot, \boxdot, \boxdot\}]} = \frac{1/6}{1/2} = \tfrac{1}{3}. \bullet \end{aligned} \tag{11.82}$$

11.4.2 Conditional Expectation

Before we give the formal definition of *conditional expectation* let us give a few motivating examples to understand it conceptually.

Let X be the temperature in Belfast. Then[7], $\mathbb{E}[X] = 13.2°$. But if we know that it's July, then $\mathbb{E}[X \mid \text{July}] = 19.3°$.

Let X be the height of a 19 y/o person. Then $\mathbb{E}[X] = 1.71\,\text{m}$. If we know that this person is from the Netherlands, then $\mathbb{E}[X \mid \text{Netherlands}] = 1.77\,\text{m}$. If, additionally, the person is male $\mathbb{E}[X \mid \text{Netherlands}, \text{Male}] = 1.83\,\text{m}$.

The expectation can be **conditional on an event**: Let $\Omega = \{1, 2, \ldots, n\}$, $\mathcal{F} = 2^\Omega$ and let P be a probability. Let X be a random variable on $(\Omega, \mathcal{F}, \mathsf{P})$. Recall that

$$\mathbb{E}[X] = \sum_{i=1}^{n} X_i \mathsf{P}[\{i\}], \tag{11.83}$$

Let $A \in \mathcal{F}$ with $\mathsf{P}[A] > 0$; then

$$\mathbb{E}[X \mid A] = \sum_{i=1}^{n} X_i \mathsf{P}[\{i\} \mid A]. \tag{11.84}$$

[7] Based on data from the Stormont Castle weather station.

Recall that

$$\mathsf{P}[\{i\} \mid A] = \frac{\mathsf{P}[\{i\} \text{ and } A]}{\mathsf{P}[A]} = \begin{cases} \frac{\mathsf{P}[\{i\}]}{\mathsf{P}[A]}, & \text{if } i \in A \\ 0, & \text{otherwise} \end{cases}, \tag{11.85}$$

therefore,

$$\mathbb{E}[X \mid A] = \tfrac{1}{\mathsf{P}[A]} \sum_{i \in A} X_i \mathsf{P}[\{i\}]. \tag{11.86}$$

Example (Die roll with insight). Consider a fair die and suppose you win £$(-1)^n n$ if you roll n: If you roll ⚀ you win −£1. If you roll ⚁ you win £2. If you roll ⚂ you win −£3, etc. Let X be your winnings. This is a random variable on $\Omega = \{1, \ldots, 6\}$. Then

$$\mathbb{E}[X] = \sum_{i=1}^{6} (-1)^i i \tfrac{1}{6} = £0.5. \tag{11.87}$$

Consider the event $A = \{⚁, ⚂, ⚃\} \subseteq \Omega$, i.e., you roll a die, someone looks at it and informs you that you rolled either ⚁ or ⚂ or ⚃. The probability of A is $\mathsf{P}[A] = \frac{1}{2}$. Then,

$$\mathbb{E}[X \mid A] = \tfrac{1}{\mathsf{P}[A]} \sum_{i \in A} X_i \mathsf{P}[\{i\}] = \frac{1}{\frac{1}{2}} \sum_{i \in A} \underbrace{(-1)^i i}_{X_i} \tfrac{1}{6} = £4. \bullet \tag{11.88}$$

Alternative formulation of the conditional expectation on an event: Let $\Omega = \{1, 2, \ldots, n\}$, $\mathcal{F} = 2^\Omega$ and P be a probability. Let $X : (\Omega, \mathcal{F}, \mathsf{P}) \to \mathbb{R}$ be a random variable with $X(\Omega) = \{X_{(1)}, X_{(2)}, \ldots, X_{(m)}\}$, i.e., X has m possible *unique* values.

Then,

$$\mathbb{E}[X] = \sum_{i=1}^{m} \mathsf{P}[X = X_{(j)}] X_{(j)}. \tag{11.89}$$

Let $A \in \mathcal{F}$ with $\mathsf{P}[A] > 0$; then

$$\begin{aligned} \mathbb{E}[X \mid A] &= \sum_{j=1}^{m} \mathsf{P}[X = X_{(j)} \mid A] X_{(j)} \\ &= \frac{1}{\mathsf{P}[A]} \sum_{j=1}^{m} \mathsf{P}[X = X_{(j)} \text{ and } A] X_{(j)}. \end{aligned} \tag{11.90}$$

The expectation of a random variable can be **conditional on another random variable**: The expectation of a random variable X on a discrete probability space can be conditioned by another random variable Y on that space. We define

$$\begin{aligned}\mathbb{E}[X \mid Y = y] &= \sum_{i=1}^{n} X_i \underbrace{\mathsf{P}[\{i\} \mid Y = y]}_{\substack{\text{Conditional on the event} \\ \{\omega \in \Omega : Y(\omega) = y\}}} \\ &= \frac{1}{\mathsf{P}[Y = y]} \sum_{i=1}^{n} X_i \mathsf{P}[\{i\} \text{ and } Y = y], \end{aligned} \tag{11.91}$$

provided that $\mathsf{P}[Y = y] > 0$.

Important: The conditional expectation, $\mathbb{E}[X \mid Y = y]$, is a function of the observation y. However, Y is a random variable; this gives rise to $\mathbb{E}[X \mid Y = Y(\omega)]$ which is a random variable — this is denoted as $\mathbb{E}[X \mid Y]$.

Moreover, $\mathbb{E}[X \mid Y = y]$ is the expectation of X conditioned on the event $A(y) = \{\omega \in \Omega : Y(\omega) = y\}$ (Caveat: this is true ONLY on discrete spaces).

11.5 Conditioning of continuous random variables

11.5.1 Conditioning on event

Definition 11.14 (Conditional expectation of random variable) *Let X be a real-valued continuous random variable on $(\Omega, \mathcal{F}, \mathsf{P})$ and $B \in \mathcal{F}$ with $\mathsf{P}[B] > 0$. We define*

$$\mathbb{E}[X \mid B] = \frac{\mathbb{E}[X 1_B]}{\mathsf{P}[B]}, \tag{11.92}$$

where 1_B is the following random variable

$$1_B(\omega) = \begin{cases} 1, & \text{if } \omega \in B \\ 0, & \text{otherwise} \end{cases} \tag{11.93}$$

provided that the expectation in Equation (11.92) exists.

Definition 11.15 (Conditional pdf) *Suppose that X is a continuous random variable with pdf p_X and $B \in \mathcal{F}$ with $\mathsf{P}[B] > 0$. We define the conditional pdf*

$$p_{X|B}(x) = \frac{1_B(x) p_X(x)}{\mathsf{P}[B]}. \tag{11.94}$$

Then it can be seen that we can determine the conditional expectation of X given an event B using the conditional $p_{X|B}$ as follows

$$\mathbb{E}[X \mid B] = \int_{-\infty}^{\infty} x p_{X|B}(x) \mathrm{d}x. \tag{11.95}$$

Given a continuous random variable $X : (\Omega, \mathcal{F}, \mathsf{P}) \to (E, \mathcal{E})$ with pdf p_X and an event $B \in \mathcal{F}$, we have that

$$\mathsf{P}[X \in A \mid B] = \int_A p_{X|B}(x) \mathrm{d}x = \int_{A \cap B} \frac{p_X(x)}{\mathsf{P}[B]} \mathrm{d}x = \frac{\mathsf{P}[X \in (A \cap B)]}{\mathsf{P}[B]}. \tag{11.96}$$

A somewhat common mistake that some people do is the following: suppose we have a real-valued random variable X with pdf p_X and we need to determine $E[X \mid X > c]$. Then it is wrong to write

$$\mathbb{E}[X \mid X > c] = \int_c^\infty p_X(x)\mathrm{d}x. \tag{11.97}$$

Instead, we need to invoke Equation (11.95). You should solve the following exercise.

Exercise 22 (Conditional expectation of normal distribution) ☕☕. Suppose that $X \sim \mathcal{N}(0,1)$. Determine $\mathbb{E}[X \mid X > 2]$, that is, determine the expectation of X if we know that $X > 2$. Hint: Use Equation (11.95). $\diamond$

11.5.2 Conditioning on random variable

Let X and Y be two real-valued (or vector-valued) continuous random variables on $(\Omega, \mathcal{F}, \mathsf{P})$, with joint pdf $p_{X,Y}(x,y)$. Let us define the conditional expectation of X conditional on the fact that $Y = y$.

Definition 11.16 (Conditional pdf and conditional expectation) *The conditional pdf of X given that $Y = y$ is defined as*

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}, \tag{11.98}$$

provided that $p_Y(y) > 0$. The conditional expectation of X given $Y = y$ is

$$\mathbb{E}[X \mid Y = y] = \int_{-\infty}^\infty x p_{X|Y}(x \mid y)\mathrm{d}x. \tag{11.99}$$

We can define the *conditional variance* of X given $Y = y$ as

$$\mathrm{Var}[X \mid Y = y] = \mathbb{E}[(X - \mathbb{E}[X \mid Y = y])^2 \mid Y = y]. \tag{11.100}$$

Note again that $\mathbb{E}[X \mid Y]$ is a random variable — and not a fixed value — and in particular it is $\mathbb{E}[X \mid Y = Y(\omega)]$.

11.5.3 Independence of events and random variables

Recall that $A, B \in \mathcal{F}$ are said to be *independent* if

$$\mathsf{P}[A \cap B] = \mathsf{P}[A]\mathsf{P}[B]. \tag{11.101}$$

We say that two random variables, X, Y, are independent if

$$\mathsf{P}[X \in A, Y \in B] = \mathsf{P}[X \in A]\mathsf{P}[Y \in B]. \tag{11.102}$$

Equivalently, two multivariate random variables X and Y with cdfs F_X and F_Y are independent if and only if

$$F_{X,Y}(x, y) = F_X(x)F_Y(y). \tag{11.103}$$

If additionally (X, Y) is continuous with pdf $p_{X,Y}$ then X and Y are independent if and only if

$$p_{X,Y}(x, y) = p_X(x)p_Y(y). \tag{11.104}$$

or, what is the same,

$$p_{X|Y}(x \mid y) = p_X(x), \text{ and } p_{Y|X}(y \mid x) = p_Y(y). \tag{11.105}$$

Additionally, for two independent real-valued random variables X and Y we can show that

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]. \tag{11.106}$$

In general this is not true if X and Y are not independent. One important consequence of the independence of two random variables, X and Y, is that

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]. \tag{11.107}$$

11.5.4 Properties

Some notable properties of the conditional expectation:

- $\mathbb{E}[X \mid Y]$ is linear: for all $a, b \in \mathbb{R}$ and RVs X_1, X_2, Y:

$$\mathbb{E}[aX_1 + bX_2 \mid Y] = a\mathbb{E}[X_1 \mid Y] + b\mathbb{E}[X_2 \mid Y]. \tag{11.108}$$

- Law of total expectation

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]. \tag{11.109}$$

- For random variables X and Y and a function g

$$\mathbb{E}[g(Y)X \mid Y] = g(Y)\mathbb{E}[X \mid Y]. \tag{11.110}$$

- If X and Y are independent,

$$\mathbb{E}[X \mid Y] = \mathbb{E}[X]. \tag{11.111}$$

- The conditional variance is given by

$$\mathrm{Var}[X \mid Y] = \mathbb{E}[X^2 \mid Y] - \mathbb{E}[X \mid Y]^2. \tag{11.112}$$

- Conditional LotUS: for a real-valued random variable X:

$$\mathbb{E}[g(X) \mid Y = y] = \int_{-\infty}^{\infty} g(x)p_{X|Y}(x \mid y)\mathrm{d}x. \tag{11.113}$$

- For two random variables X, Y (not necessarily independent), and a function g

$$\mathbb{E}[g(X,Y) \mid Y = y] = \mathbb{E}[g(X,y) \mid Y = y]. \tag{11.114}$$

- Law of total expectation (again). For random variables X and Y and a function g,

$$\mathbb{E}\Big[\mathbb{E}[g(X,Y) \mid Y]\Big] = \mathbb{E}[g(X,Y)]. \tag{11.115}$$

The law of total expectation is particularly useful and allows us to determine the expectation of a random variable X in cases where it is easier to determine the conditional expectation of X given a random variable Y. Let us give an example.

Example (Expectation of sum of random length). Suppose that the expected amount of money a customer spends at a store is £10. Every day, the expected number of customers is 50. What is the expected income of the store?

Let X_i be the amount of money that the i-th client will spend and $i = 1, \ldots, N$. Note that all X_i for $i = 1, \ldots, N$ and N are random variables. The total amount of money that the store earns is $Y = X_1 + \ldots + X_N$.

Assume that N is known. Then, $\mathbb{E}[Y \mid N] = \mathbb{E}[X_1 + \ldots + X_N \mid N] = 10N$. Now using the law of total expectation we have $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid N]] = \mathbb{E}[10N] = £500$. •

Exercise (Law of total variance) ☕☕. Similar to the law of total expectation, we have the law of total variance which states that for a real-valued random variable and a random variable Y we have

$$\text{Var}[X] = \mathbb{E}[\text{Var}[X \mid Y]] + \text{Var}[\mathbb{E}[X \mid Y]]. \tag{11.116}$$

To prove this follow these steps:

1. Apply the expectation to both sides of Equation (11.112)
2. apply the law of total expectation to the term $\mathbb{E}[\mathbb{E}[X^2 \mid Y]]$ that appears in Step 1
3. Use Definition 11.8 to determine the variance of $\mathbb{E}[X \mid Y]$ and
4. apply the law of total expectation to the term $\mathbb{E}[\mathbb{E}[X \mid Y]]$ that appears in Step 3

Lastly, combine the above equations to arrive at Equation (11.116). ⋄

Example (Application of the law of total variance). Following up on the above example, suppose that now we have $\mathbb{E}[X_i] = 10$, $\text{Var}[X_i] = £^2100$ and $\mathbb{E}[N] = 50$, $\text{Var}[N] = 20$. Let us determine the variance of $Y = X_1 + \ldots + X_N$. By the law of total variance given in Equation (11.116) we have that

$$\text{Var}[Y] = \mathbb{E}[\text{Var}[Y \mid N]] + \text{Var}[\underbrace{\mathbb{E}[Y \mid N]}_{10N}]. \tag{11.117}$$

We now need to determine $\mathbb{E}[\text{Var}[Y \mid N]]$. We have that $\text{Var}[Y \mid N]$ is given by

$$\text{Var}[Y \mid N] = \text{Var}[X_1 + \ldots + X_N \mid N] \tag{11.118}$$

Under the assumption that $X_1, \ldots, X_N$ are mutually independent we have

$$= \text{Var}[X_1 \mid N] + \text{Var}[X_2 \mid N] + \ldots + \text{Var}[X_N \mid N] \tag{11.119}$$

and assuming that X_i and N are independent for all i

$$= 20N, \tag{11.120}$$

so from Equation (11.117) we have

$$\mathrm{Var}[Y] = \mathbb{E}[20N] + \mathrm{Var}[10N] = 20\mathbb{E}[N] + 100\,\mathrm{Var}[N] = 3000. \bullet \tag{11.121}$$

Exercise 23 (Optimality of conditional expectation) ☕☕☕**.** (i) Given two vector-valued n-dimensional jointly distributed random variables X and Y, define the function $f : \mathbb{R}^n \to \mathbb{R}$

$$f(z; y) = \mathbb{E}[\|X - z\|^2 \mid Y = y], \tag{11.122}$$

where $z = z(y)$ is any estimate of X given $Y = y$ (not necessarily the conditional expectation). Then $f(z) \geq f(\mathbb{E}[X \mid Y = y])$ for all $z \in \mathbb{R}^n$. In other words, $\mathbb{E}[X \mid Y = y] \in \arg\min_z f(z; y)$. (ii) State and prove the counterpart of Exercise 21 for the conditional expectation. ⋄

11.5.5 Conditioning of multivariate normals

Theorem 11.17 (Conditioning of multivariate normal) *Let $X \sim \mathcal{N}(\mu, \Sigma)$ be an n-dimensional random vector. Let us partition X into two random vectors X_1 and X_2 as follows*

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \tag{11.123}$$

with $X_1 \in \mathbb{R}^{n_1}$, $X_2 \in \mathbb{R}^{n_2}$ with $n = n_1 + n_2$. Let

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \text{ and } \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \tag{11.124}$$

and assume that $\Sigma_{22} \in \mathbb{S}^{n_2}_{++}$. Then, the conditional distribution of X_1 given that $X_2 = x_2$ is normal with mean

$$\mathbb{E}[X_1 \mid X_2 = x_2] = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \tag{11.125}$$

and

$$\mathrm{Var}[X_1 \mid X_2 = x_2] = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{11.126}$$

Proof. The proof hinges on *Schur's complement*. We define the Schur complement of Σ (with respect to Σ_{22}) to be the following nonsingular matrix

$$\Sigma_* = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{11.127}$$

Then, the inverse of Σ is the matrix

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_{11}^* & \Sigma_{12}^* \\ \Sigma_{21}^* & \Sigma_{22}^* \end{bmatrix}, \tag{11.128}$$

where $\Sigma_{11}^* = \Sigma_*^{-1}$, $\Sigma_{12}^* = -\Sigma_*^{-1}\Sigma_{12}\Sigma_{22}^{-1}$, $\Sigma_{21}^* = (\Sigma_{12}^*)^\intercal$ (since Σ is symmetric) and $\Sigma_{22}^* = \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}\Sigma_*^{-1}\Sigma_{12}\Sigma_{22}^{-1}$. We need to determine the pdf of X_1 conditional on X_2; by Equation (11.98) we have that $p_{X_1|X_2}(x_1 \mid x_2)$ is *proportional to* $p_{X_1,X_2}(x_1, x_2)$; note that the denominator of Equation (11.98) is $p_{X_2}(x_2)$, which independent of x_1. We denote this as

$$p_{X_1|X_2}(x_1 \mid x_2) \propto p_{X_1,X_2}(x_1, x_2). \tag{11.129}$$

Since (X_1, X_2) are jointly normal, we have that its pdf is (see Equation (11.71))

$$p_{X_1,X_2}(x_1, x_2) \propto \exp\left(-\tfrac{1}{2}(x-\mu)^\intercal\Sigma^{-1}(x-\mu)\right), \tag{11.130}$$

where $x = (x_1, x_2)$ and $\mu = (\mu_1, \mu_2)$. The reader can use the block-inversion formula in Equation (11.128) to verify that we can write

$$(x-\mu)^\intercal\Sigma^{-1}(x-\mu) = (x_1-\mu_*)^\intercal\Sigma_*^{-1}(x_1-\mu_*) + (x_2-\mu_2)^\intercal\Sigma_{22}^{-1}(x_2-\mu_2), \tag{11.131}$$

where $\mu_* = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)$. From Equations (11.129) and (11.130) we conclude that

$$p_{X_1|X_2}(x_1 \mid x_2) \propto \exp\left(-\tfrac{1}{2}(x_1-\mu_*)^\intercal\Sigma_*^{-1}(x_1-\mu_*)\right), \tag{11.132}$$

which proves that $X_1 \mid X_2$ is normal with mean μ_* and variance Σ_*. □

Remark. By Equation (11.126), we have

$$\mathrm{Var}[X_1 \mid X_2 = x_2] = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{11.133}$$

Since $\Sigma_{22} \succ 0$, $\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \succcurlyeq 0$, therefore $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \preccurlyeq \Sigma_{11}$, i.e.,

$$\mathrm{Var}[X_1 \mid X_2 = x_2] \preccurlyeq \mathrm{Var}[X_1]. \tag{11.134}$$

In other words, additional information *does not "increase" (in the sense of $\preccurlyeq$) the uncertainty!*

Example (Conditioning of normally distributed random variables). Suppose that $Z = (Z_1, Z_2, Z_3, Z_4)$ is a four-dimensional random variable that follow the normal distribution, $Z \sim \mathcal{N}(\mu, \Sigma)$ with $\mu = (1, 2, 3, 4)$ and

$$\Sigma = \begin{bmatrix} 1.0 & 0.35 & 0.32 & 0.39 \\ 0.35 & 0.84 & 0.3 & 0.26 \\ 0.32 & 0.3 & 0.77 & 0.23 \\ 0.39 & 0.26 & 0.23 & 0.83 \end{bmatrix}. \tag{11.135}$$

The reader can verify that $\Sigma \in \mathbb{S}^4_{++}$. Suppose we measure Z_3 and Z_4 and we want to determine $\mathbb{E}[Z_1, Z_2 \mid Z_3, Z_4]$. We will apply Theorem 11.17 with $X_1 = (Z_1, Z_2)$ and $X_2 = (Z_3, Z_4)$. We have

$$\Sigma_{11} = \begin{bmatrix} 1.0 & 0.35 \\ 0.35 & 0.84 \end{bmatrix}, \Sigma_{12} = \begin{bmatrix} 0.32 & 0.39 \\ 0.3 & 0.26 \end{bmatrix}, \Sigma_{22} = \begin{bmatrix} 0.77 & 0.23 \\ 0.23 & 0.83 \end{bmatrix}. \tag{11.136}$$

and $\mu_1 = (1, 2)$, $\mu_2 = (3, 4)$. By Theorem 11.17

$$\mathbb{E}\left[Z_1, Z_2 \,\middle|\, \begin{matrix} Z_3 = z_3 \\ Z_4 = z_4 \end{matrix} \right] = \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 0.32 & 0.39 \\ 0.3 & 0.26 \end{bmatrix} \begin{bmatrix} 0.77 & 0.23 \\ 0.23 & 0.83 \end{bmatrix}^{-1} \left(\begin{bmatrix} z_3 \\ z_4 \end{bmatrix} - \begin{bmatrix} 3 \\ 4 \end{bmatrix} \right), \tag{11.137}$$

and

$$\mathrm{Var}\left[Z_1, Z_2 \,\middle|\, \begin{matrix} Z_3 = z_3 \\ Z_4 = z_4 \end{matrix} \right] = \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 0.32 & 0.39 \\ 0.3 & 0.26 \end{bmatrix} \begin{bmatrix} 0.77 & 0.23 \\ 0.23 & 0.83 \end{bmatrix}^{-1} \begin{bmatrix} 0.32 & 0.3 \\ 0.39 & 0.26 \end{bmatrix}. \bullet \tag{11.138}$$

Exercise 24 (Conditioning of normals) ☕. Let Z be as in the example above and we measure $Z_2 = 2.5$, $Z_3 = 2.8$, and $Z_4 = 4.1$. Determine the conditional expectation of Z_1 given these measurements. ⋄

Exercise 25 (Conditioning of normals) ☕☕. Let Z be as in the example above and we measure $Z_2 = 2.5$, and $Z_4 = 4.1$. Determine the conditional expectation and the conditional variance of (Z_1, Z_3) given these measurements. ⋄