Probability Notes

Amro Al Baali

Compiled: November 4, 2020

Contents

Preface

Most of the notes in Part I are based off ECSE 509: "*Probability and Random Signals 2*" course taught at McGill University in Fall 2019 by Prof. I. Psaromiligkos [1].

Notations

Notation	Definition
S	Sample space.
$\mathcal{P}$	Power set.
$\mathcal{F}$	Events algebra.
$P : \mathcal{P} \to [0, 1]$	Probability function.
$f : \mathbb{R}^n \to \mathbb{R}$	Probability density function (PDF).
$F : \mathbb{R}^n \to \mathbb{R}$	Cumulative distribution function (CDF).
$p : S \subset \mathbb{R}^n \to \mathbb{R}$	Cumulative distribution function (CDF). Usually used for discrete random variables.
$\underline{(\cdot)} : S \to \mathbb{R}^n$	A random variable (RV). S is referred to as the *domain*.
$\mathcal{R}_x$	Set of possible points that random variable $\underline{x}$ can take.
$\underline{x}$	Single random variable.
$\underline{\mathbf{x}}$	Multivariate random variable.
$\mathbb{E}[\cdot]$	Expectation operator.
$\mathrm{Var}[\cdot]$	Variance.
$\jmath$	$\sqrt{-1}$.

Part I

Fundamentals

Chapter 1

Single Variable Probability

1.1 Definitions and intuition

Probability is a measure of:

- Our certainty or belief that a statement is true. My belief could be different than yours;
- How frequently an event will occur.

Definition 1.1.1 (*Random experiment*)**.** A *random experiment* is an experiment (or a physical process) whose outcome is not certain but all of its possible outcomes are known and predictable in advance.

A random experiment is modeled as a probability space (check Definition 1.1.7).

Definition 1.1.2 (*Sample space S*)**.** The *sample space* S is the set of *all* possible *outcomes*.

Definition 1.1.3 (*Event*)**.** An *event* A is a *subset* of S. That is, $A \subseteq S$. It's NOT an element of S! It is rather an element of the power set $\mathcal{P}_S$ (Definition 1.1.4).

Thus, an event A is a *set* of outcomes.

Definition 1.1.4 (*Power set of S*)**.** The *power set* of the sample space S is the set containing all subsets of a set S. It's denoted by $\mathcal{P}_S$, or simply $\mathcal{P}$.

- $\mathcal{P}_S$ is the set of all events.
- $\varnothing \subseteq S$: Impossible event.
- $S \subseteq S$: Certain event.

Example 1.1.1 (Flipping a coin)**.** Flipping a coin *once* has the sample space

$$S = \{H, T\}, \tag{1.1}$$

H and T are all the possible *outcomes* (NOT events) of the sample space S. The power set is thus given by

$$\mathcal{P}_S = \{\{\varnothing\}, \{H\}, \{T\}, \{H, T\}\}. \tag{1.2}$$

Note that the last element of $\mathcal{P}_S$, $\{H, T\}$ is the sample space and it is thus a *certain* event. △

Example 1.1.2 (Number of occurences)**.** Relative frequency interpretation of probability (which is very intuitive). Say a random experiment gives an outcome a in an event A (i.e., $a \in A \subseteq S$) from the sample space. Repeat the experiment N times. Then, define the number of occurrences of event A after repeating the experiment N times by

$$n(A, N) = \#\text{ of times that } A \text{ happens}. \tag{1.3}$$

Note that "# of times that A happens" implies that the outcome a_k of experiment k belongs to the event A. That is, $a_k \in A$. Then,

$$\underbrace{\lim_{N\to\infty} \frac{n(A, N)}{N}}_{\text{Probability of event } A.} = \text{constant}. \tag{1.4}$$

△

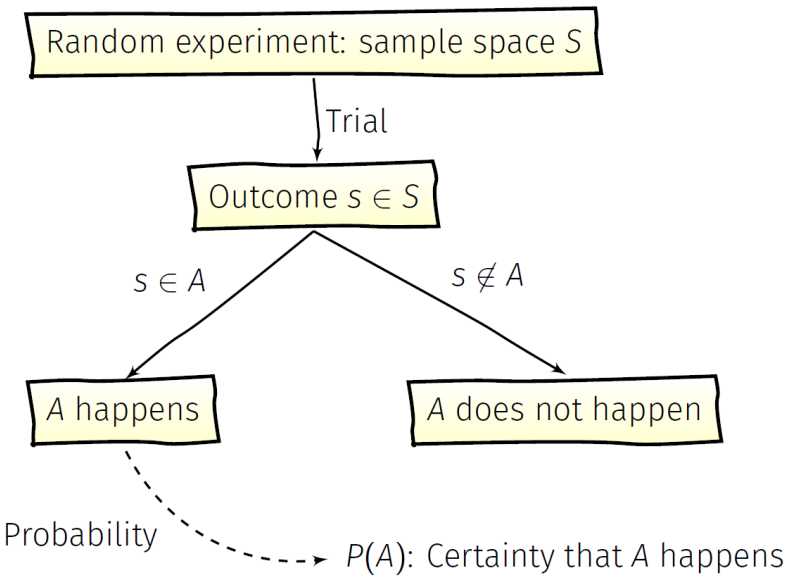

Figure 1.1: Intuition of probability. Figure obtained from [1].

Definition 1.1.5 (*Probability*)**.** *Probability* is a set function $P : \mathcal{P}_S \mapsto [0, 1]$ that assigns to an event

$A \in \mathcal{P}_S$ (recall $A \subseteq S$) a number $P(A)$ that satisfies the following three axioms:

1. $P(A) \geq 0$,
2. $P(S) = 1$, and
3. For any sequence of mutually exclusive events $A_1, A_2, \ldots$ (i.e., for which $A_i \bigcap A_j = \emptyset$ for $i \neq j$), we have

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \tag{1.5}$$

The number $P(A)$ is called the probability of A.

Remark 1.1.1. It is NOT always possible to define a function $P : \mathcal{P} \mapsto [0, 1]$ that satisfies the three axioms. Workaround: focus on the events in a σ-algebra $\mathcal{F}$.

Here are some other remarks about the probability function.

- The probability function $P : \mathcal{P}_S \mapsto [0, 1]$ maps the power set $\mathcal{P}_S$ (or σ-algebra) to $[0, 1]$. That is, the argument of P should be an *element* of $\mathcal{P}_S$. And since $\mathcal{P}_S$ is a *set of sets* of S, then the argument is the event which is denoted by $\{a_1, a_2, \ldots, a_n\} = A \in \mathcal{P}_S$, where $a_1, \ldots, a_n \in S$. Thus the proper way to write "the probability of event A" is $P(\{A\})$ but the notation $P(A)$ will often be used for ease of reading and writing.
- Note that $P(A) = 0$ implies that event A is *improbable*; not necessarily impossible! It's still possible to get A. This might be a little unintuitive, but it can be thought of that the chance of A happening is so small that it seems that event A is highly unlikely to occur.
- Similarly, $P(A) = 1$ does NOT imply that A will certainly occur! It just implies that it is very probably.

Definition 1.1.6 (σ-*algebra*)**.** A set $\mathcal{F}$ of subsets of S, that is, $\mathcal{F} \subseteq \mathcal{P}$ is σ-*algebra* if and only if

1. $S \in \mathcal{F}$,
2. $E \in \mathcal{F} \implies E^{\mathrm{C}} \in \mathcal{F}$, and
3. If the sets $A_1, A_2, \ldots$ belong to $\mathcal{F}$, then so does $\bigcup_{i=1}^{\infty} A_i$.

If S is countable, then $\mathcal{F} = \mathcal{P}_S$ is used.

A random experiment is modeled as a probability space.

Definition 1.1.7 (*Probability space*)**.** A *probability space* is a triplet

$$(S, \mathcal{F}, P), \tag{1.6}$$

where

1. S: is the sample set,
2. $\mathcal{F}$ is the events algebra, and
3. $P : \mathcal{F} \mapsto [0, 1]$ is the probability function.

1.2 Basic theorems

Theorem 1.2.1 (*Basic theorems*)**.** Below is a list of basic theorems of probability.

1. For any event $A \in \mathcal{F}$,
$$P\left(A^{\mathrm{C}}\right) = 1 - P(A). \tag{1.7}$$
2. For any event $A \in \mathcal{F}$,
$$0 \leq P(A) \leq 1. \tag{1.8}$$
3. $P(\emptyset) = 0$.
4. For any events $A, B \in \mathcal{F}$,
$$P(A - B) = P(A) - P(A \cap B). \tag{1.9}$$
Special case: if $B \subseteq A$, then
$$P(A - B) = P(A) - P(B). \tag{1.10}$$
5. For any $A, B \in \mathcal{F}$,
$$P(A) = P(AB) + P\left(AB^{\mathrm{C}}\right). \tag{1.11}$$
6. For any $A, B \in \mathcal{F}$,
$$P(A \cap B) = P(A) + P(B) - P(A \cup B), \tag{1.12}$$
$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{1.13}$$

1.3 Conditional probabilities

Events can affect other events. Therefore, the question "what's the probability of event A happening?" could be quite different from "what's the probability of event A *given that event* B *happened*?"[1]. The notation

$$P\left(A|\ B\right) \tag{1.14}$$

reads "probability of event A happening given that event B happened".

Given

1. a random experiment $(S, \mathcal{F}, P\left(\right))$, and
2. $A,\ B \in \mathcal{F}$ with $P\left(B\right) \neq 0$.

Definition 1.3.1 (*Conditional probability*)**.** The conditional probability of A given B, $P\left(A|\ B\right)$, defined as

$$P\left(A|B\right) = \frac{P\left(A \bigcap B\right)}{P\left(B\right)}. \tag{1.15}$$

Theorem 1.3.1 (Conditional probability)**.** Let $B \in \mathcal{F}$ with $P\left(B\right) \neq 0$. The mapping

$$P\left(\cdot|B\right) : \mathcal{F} \to \mathbb{R}, \tag{1.16}$$

satisfies the axioms of probability

1. $P\left(A|B\right) \geq 0, \quad \forall A \in \mathcal{F}$
2. $P\left(S|B\right) = 1$
3. If $A_1, A_2, \ldots \in \mathcal{F}$ is a sequence of mutually exclusive events, then

$$P\left(\bigcup_{i=1}^{\infty} A_i|\ B\right) = \sum_{i=1}^{\infty} P\left(A_i|B\right). \tag{1.17}$$

Therefore,

- For a given event B with $P\left(B\right) \neq 0$, the mapping

$$P\left(\cdot|B\right) : \mathcal{F} \to \mathbb{R}, \tag{1.18}$$

  defines a valid probability function.

- All the basic theorems of probability listed in Theorem 1.2.1apply to $P\left(\cdot|B\right)$.

[1] In some cases, the two questions are the same. This brings the notion of independence.

Here are some properties of conditional probability

1. For $A, B \in \mathcal{F}$ with $P(A), P(B) \neq 0$:

$$P(AB) = P(A|B)\,P(B) \tag{1.19}$$
$$= P(B|A)\,P(A). \tag{1.20}$$

2. **Total probability**. Let $B_1, B_2, \ldots, B_n$ be a partition of S, with $P(B_i) \neq 0, \forall i$:

$$P(A) = P(A|B_1)\,P(B_1) + P(A|B_2)\,P(B_2) + \ldots + P(A|B_n)\,P(B_n) \tag{1.21}$$
$$= \sum_{i=1}^{n} P(A|B_i)\,P(B_i). \tag{1.22}$$

3. **Bayes rule**. Let $B_1, B_2, \ldots, B_n$ be a partition of S, with $P(B_i) \neq 0, \forall i$:

$$P(B_i|A) = \frac{P(A|B_i)\,P(B_i)}{\sum_{j=1}^{n} P(A|B_j)\,P(B_j)}. \tag{1.23}$$

Special case for $n = 1$:

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}. \tag{1.24}$$

Definition 1.3.2 (*Indpendence*)**.** The events $A, B \in \mathcal{F}$ are called *independent* if

$$\underbrace{P(A|B) = P(A)}_{\text{requires } P(B)\neq 0} \text{ or } \underbrace{P(B|A) = P(B)}_{\text{requires } P(A)\neq 0} \text{ or } P(AB) = P(A)\,P(B) \tag{1.25}$$

Definition 1.3.3 (*Independence (multiple events)*)**.** The events $A, B, C \in \mathcal{F}$ are called *independent* if

- They are independent in pairs, and
- $P(ABC) = P(A)\,P(B)\,P(C)$.

Definition 1.3.4 (*Conditional independence*)**.** Events $A, B \in \mathcal{F}$ are called *conditionally independent given $C \in \mathcal{F}$, with $P(C) \neq 0$*, if

$$P(AB|C) = P(A|C)\,P(B|C). \tag{1.26}$$

Remark: Conditional independence is quite different from independence. They do not imply each other. Here are some examples.

Example 1.3.1 (Independent but conditionally dependent)**.** Two fair coins are flipped. Define the fol-

lowing events

- A - Your first coin flip is heads,
- B - Your second coin flip is heads, and
- C - Your first two flips were the same.

Then A and B are independent. However, A and B are conditionally dependent given C, since if you know C then your first coin flip will inform the other one.

Example 1.3.2 (Dependent but conditionally independent)**.** Consider two brothers John and Joseph, both having a genetic disease. These two events are dependent as they are brothers. However, given the condition that Joseph is an adopted son of the family makes the events independent.

1.4 Random variables and distributions

In many random experiments, it is convenient to assign numerical labels to outcomes.

Example 1.4.1. Experiment: pick a car out of a black Escort (Eb), a red Escort (Er), and a red Mazda (Mr). The sample space is then

$$S = \{Eb, Er, Mr\}, \tag{1.27}$$

and the events algebra is $\mathcal{F} = \mathcal{P}$.

Say we are interested in the make of the car and not the color. Then we can assign labels as can be seen in Figure 1.2.

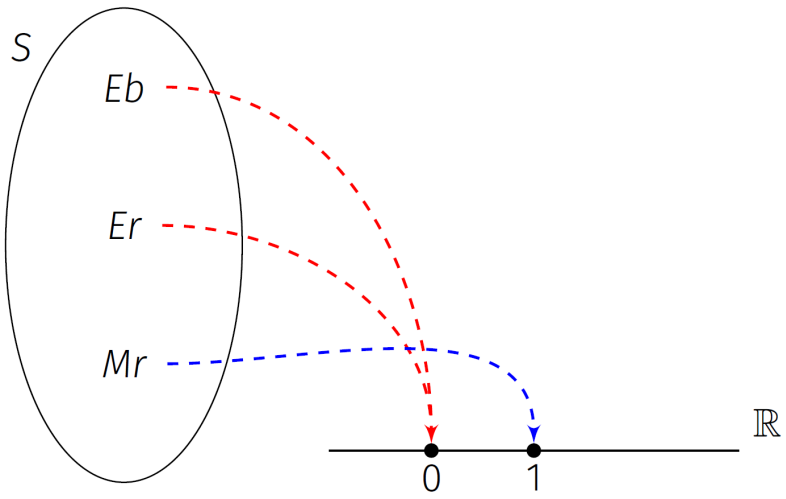

Figure 1.2: Assigning numerical labels to outcomes; each color is an outcome. From [1]

Definition 1.4.1 (*Random variable (RV)*)**.** A *random variable (RV)* (denoted by an underline) $\underline{x} : S \to \mathbb{R}$ is a mapping from S to $\mathbb{R}$ with the following properties.

1. $A(x) = \{s \in S : \underline{x}(s) \leq x\} \in \mathcal{F}, \forall x \in \mathbb{R}$
2. $P(\{s \in S : \underline{x}(s) = \infty\}) = 0$ and $P(\{s \in S : \underline{x}(s) = -\infty\}) = 0$

S is referred to as the *domain* of the RV $\underline{x}(\cdot)$ (since a RV is a mapping, then it has a domain and a range). [The domain can be thought of as the sample space[a].]

[a] Not quite sure of this statement.

Definition 1.4.2 (*Cumulative distribution function (CDF)*)**.** The function $F : \mathbb{R} \to [0, 1]$ defined by

$$F(x) = P(s \in S : \underline{x}(s) \leq x) \tag{1.28}$$
$$= P(\underline{x} \leq x), \quad \forall x \in \mathbb{R} \tag{1.29}$$

is called the *cumulative distribution function (CDF)* of $\underline{x}$. Note that (1.29) is a shorthand for (1.28).

Note: sometimes the CDF will be denoted with a subscript of the RV (e.g., $F_x(x)$). This is for clarity as to make sure that the CDF is over the RV $\underline{x}$. This come more in handy when there are multiple variables in hands (such as in integration).

1.4.1 Properties of random variables

Here are some properties of random variables.

1. $\lim_{x\to\infty} F_x(x) = 1$ and $\lim_{x\to-\infty} F_x(x) = 0$
2. $F_x(x)$ is non-decreasing
3. $F_x(x)$ is continuous from the right. That is

$$\lim_{x\to x_0^+} F_x(x) = F_x(x_0) \tag{1.30}$$

4. $P(\{x_1 < \underline{x} \leq x_2\}) = F_x(x_2) - F_x(x_1)$
5. $P(\{\underline{x} = x_1\}) = F_x(x_1) - F_x(x_1^-)$, where $F_x(x_1^-) := \lim_{x\to x_1^-} F_x(x)$
6. $P(\{x_1 \leq \underline{x} \leq x_2\}) = F_x(x_2) - F_x(x_1^-)$

Definition 1.4.3 (*Types of RVs*)**.** A RV $\underline{x}$ is called

- *Discrete* if $F_x(x)$ is constant except of countable number of discontinuities (piecewise constant). Figure 1.3 shows an example of a CDF of a discrete RV.
- *Continuous* if $F_x(x)$ is
  1. continuous, and
  2. differentiable (with the exception of a countable number of point).
- *Mixed* if it is neither continuous nor discrete.

Definition 1.4.4 (*Probability mass function (PMF)*)**.** Let $\underline{x}$ be a discrete RV. Then,

- $P(\underline{x} = x) = 0$ when $x \neq x_i, i = 1, 2, \ldots,$
- $\mathcal{R}_x = \{x_i : i = 1, , 2, \ldots\}$ is the set of possible points.

The *probability mass function (PMF)* of a RV $\underline{x}$, $p(x)$ or $p_x(x)$, $x \in \mathbb{R}$, is defined as

$$p(x) = P(\underline{x} = x) \qquad = \begin{cases} 0 & \text{if } x \notin \mathcal{R}_x, \\ P(\underline{x} = x_i) & \text{if } x \in \mathcal{R}_x, \text{ i.e., } x = x_i, i = 1, 2, \ldots. \end{cases} \tag{1.31}$$

Also,

- $\sum_{i=1}^{\infty} p(x_i) = 1$ and $p(x) \geq 0, \forall x$ (defining properties),
- $P(\underline{x} \in \mathcal{A}) = \sum_{x_i \in \mathcal{A}} p(x_i)$.

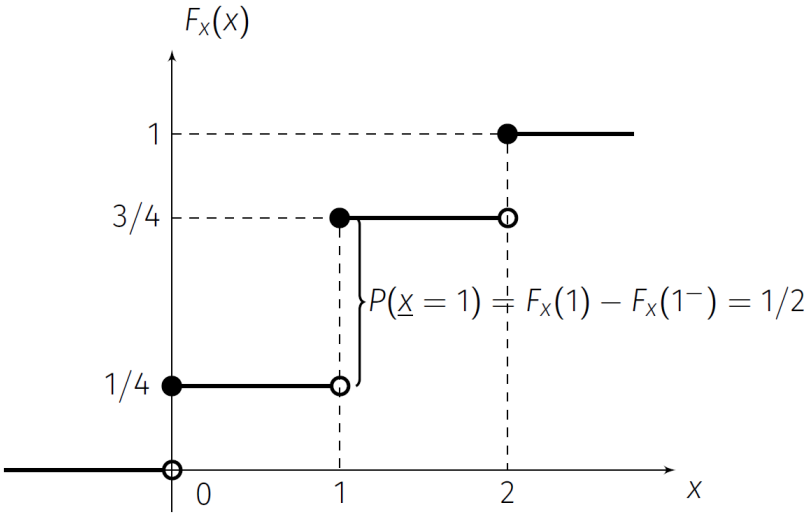

Figure 1.3: Example of a CDF of a discrete RV. From [1].

Definition 1.4.5 (*Porbability density function (PDF)*)**.** The *probability density function (PDF)*, $f(x)$ or $f_x(x)$, of a RV $\underline{x}$ is defined as

$$f_x(x) = \frac{\mathrm{d}F_x(x)}{\mathrm{d}x}. \tag{1.32}$$

1.4.2 Properties of PDFs

If $\underline{x}$ is continuous with PDF $f(x)$

1. $f_x(\underline{x}) \geq 0, \forall x$ (defining property)
2. $\int_{-\infty}^{x} f_x(\lambda)\mathrm{d}\lambda = F_x(x)$
3. $\int_{-\infty}^{\infty} f_x(x)\mathrm{d}\lambda = 1$ (defining property)

4.

$$P\left(\{a \leq \underline{x} \leq b\}\right) = \int_a^b f_x(\lambda)\mathrm{d}\lambda \tag{1.33}$$

$$P\left(\{a < \underline{x} \leq b\}\right) = P\left(\{a \leq \underline{x} < b\}\right) = P\left(\{a < \underline{x} < b\}\right). \tag{1.34}$$

Interpretation of PDF: For a continuous $\underline{x}$: $P\left(\underline{x} = x\right) = 0$ for all $x \in \mathbb{R}$ which implies that the PDF $f\left(x\right)$ is *not* a probability.

For small $\epsilon > 0$:

$$P\left(|\underline{x} - x| < \epsilon\right) = P\left(x - \epsilon < \underline{x} < x + \epsilon\right) \tag{1.35}$$

$$= \int_{x-\epsilon}^{x+\epsilon} f\left(t\right)\mathrm{d}t \tag{1.36}$$

$$\approx 2\epsilon f\left(x\right) \tag{1.37}$$

assuming $f\left(x\right)$ is continuous at x. Equivalently,

$$f\left(x\right) \approx \frac{1}{2\epsilon} P\left(x - \epsilon < \underline{x} < x + \epsilon\right). \tag{1.38}$$

$f\left(x\right)$ is proportional to the probability that $\underline{x}$ lies in a small neighbourhood of x.

1.4.3 Conditional distributions

Definition 1.4.6 (*Conditional distributions*)**.** Let $M \in \mathcal{F}$ with $P\left(M\right) \neq 0$. Then

1. *Conditional CDF* of $\underline{x}$ given event M is

$$F_x(x|M) = P\left(\underline{x} \leq x|M\right) \tag{1.39}$$

$$= \frac{P\left(\underline{x} \leq x, M\right)}{P\left(M\right)}. \tag{1.40}$$

2. *Conditional PDF* of $\underline{x}$ given event M is

$$f_x(x|M) = \frac{\mathrm{d}F_x(x|M)}{\mathrm{d}x}. \tag{1.41}$$

3. *Conditional PMF* of $\underline{x}$ ($\underline{x}$ discrete given M) is

$$p\left(x|M\right) = \begin{cases} 0, & x \notin \{x_i : i = 1, 2, \ldots\} \\ P\left(\underline{x} = x_i|M\right), & x \in \{x_i : i = 1, 2, \ldots\}. \end{cases} \tag{1.42}$$

Note that conditional CDFs, PDFs, and PMFs behave like unconditional ones.

Properties of conditional distributions

Let $B_1, B_2, \ldots, B_n$ be a partition of S, $P(B_i) \neq 0$. Let $A \in \mathcal{F}$, $P(A) \neq 0$.

- **Total probability**:

$$F_x(x) = \sum_{i=1}^{n} F_x(x|B_i) P(B_i), \tag{1.43}$$

$$f_x(x) = \sum_{i=1}^{n} f_x(x|B_i) P(B_i), \tag{1.44}$$

$$p_x(x) = \sum_{i=1}^{n} p_x(x|B_i) P(B_i). \tag{1.45}$$

- **Bayes rule**:

$$P(A|\underline{x} \leq x) = \frac{P(\underline{x} \leq x, A)}{P(\underline{x} \leq x)} \tag{1.46}$$

$$= \frac{P(\underline{x} \leq x|A) P(A)}{P(\underline{x} \leq x)} \tag{1.47}$$

$$= \frac{F_x(\underline{x} \leq x|A) P(A)}{F_x(\underline{x} \leq x)}. \tag{1.48}$$

When $\underline{x}$ is discrete with $\mathcal{R}_x = \{x_1, x_2, \ldots\}$, $p(x_i) \neq 0$:

- **Conditional probability** using PMF:

$$P(A|\underline{x} = x_i) = \frac{P(A, \underline{x} = x_i)}{P(\underline{x} = x_i)} \tag{1.49}$$

$$= \frac{P(A, \underline{x} = x_i)}{p_x(\underline{x} = x_i)}. \tag{1.50}$$

- **Total probability**:

$$P(A) = \sum_{i=1}^{\infty} P(A|\underline{x} = x_i) p(x_i). \tag{1.51}$$

- **Bayes rule**:

$$p(x|A) = \frac{P(A|\underline{x} = x) p(x)}{P(A)} \tag{1.52}$$

$$= \frac{P(A|\underline{x} = x) p(x)}{\sum_{i=1}^{\infty} P(A|\underline{x} = x_i) P(x_i)}. \tag{1.53}$$

When $\underline{x}$ is continuous, it is not possible to divide by $P(\underline{x} = x) = 0$. Therefore, another definition for conditional PDF is needed.

Definition 1.4.7 (*Conditional PDF*)**.** The *conditional PDF* of $A \in \mathcal{F}$ given $\underline{x} = x$ (assuming $f_x(x) \neq 0$) is

$$P\left(A|\underline{x} = x\right) = \frac{f_x(x|A)P\left(A\right)}{f_x(x)}. \tag{1.54}$$

The properties of such conditional PDF are

- **Total probability**: $P\left(A\right) = \int_{-\infty}^{\infty} P\left(A|\underline{x} = x\right) f_x(x)\mathrm{d}x$*/*
- **Bayes rule**:

$$f_x(x|A) = \frac{P\left(A|\underline{x} = x\right) f_x(x)}{P\left(A\right)} \tag{1.55}$$

$$= \frac{P\left(A|\underline{x} = x\right) f_x(x)}{\int_{-\infty}^{\infty} P\left(A|\underline{x} = x\right) f_x(x)\mathrm{d}x}. \tag{1.56}$$

1.5 Transformed random variables

Say there exists a random variable $\underline{x} : S \to \mathbb{R}$ that is applied on the probability space $(S, \mathcal{F}, P)$. Further, say there exists a [deterministic] function $g : \mathbb{R} \to \mathbb{R}$. Then, there is a new RV $\underline{y}$ defined as

$$\underline{y}(s) = g(\underline{x}(s)) \quad \forall s \in S. \tag{1.57}$$

Furthermore, say the distribution (i.e., CDF and PDF (or PMF)) of $\underline{x}$ is known. THen what is the distribution of $\underline{y} = g(\underline{x})$?

The foundation is the following

$$P\left(\underline{y} \in A\right) = P\left(\left\{s \in S : \underline{y}(s) \in A\right\}\right) \tag{1.58}$$

$$= P\left(\left\{s \in S : g(\underline{x}(s)) \in A\right\}\right). \tag{1.59}$$

There are different methods to find the distribution of a transformed random variable (i.e., $\underline{y} = g(\underline{x})$) depending on the types of variables.

1. **Method of distributions** works when for *any* types of random variables (continuous or discrete). However, it is the most exhaustive method.
2. **Method of transformations** works for *continuous* random variables; i.e., both $\underline{x}$ and $\underline{y}$ are continuous.
3. **Discrete transformation** for *discrete* transformed variable $\underline{y}$ and *continuous* or *discrete* domain random variable $\underline{x}$.

1.5.1 Method of distribution

The method will be motivated by an example.

Example 1.5.1. Consider the transformed random variable

$$\underline{y} = g(\underline{x}) \tag{1.60}$$
$$= \underline{x}^2. \tag{1.61}$$

Evaluate the CDF as follows

$$F_y(y) = P\left(\underline{y} \leq y\right) = P\left(g(\underline{x}) \leq y\right) = P\left(\underline{x}^2 \leq y\right) \tag{1.62}$$
$$= P\left(\left\{s \in S : \underline{x}(s)^2 \leq y\right\}\right). \tag{1.63}$$

There are three cases

1. $y < 0$: $P\left(\underline{x}^2 \leq y\right) = 0 \implies F_y(y) = 0$.
2. $y = 0$: $P\left(\underline{x}^2 \leq 0\right) = P\left(\underline{x}^2 = 0\right) = P\left(\underline{x} = 0\right) \implies F_y(0) = P\left(\underline{x} = 0\right)$.
3. $y > 0$: $P\left(\underline{x}^2 \leq y\right) = P\left(|\underline{x}| \leq \sqrt{y}\right) = P\left(-\sqrt{y} \leq \underline{x} \leq \sqrt{y}\right) \implies$

$$F_y(y) = \int_{-\sqrt{y}}^{\sqrt{y}} f_x(x)\mathrm{d}x. \tag{1.64}$$

1.5.2 Method of transformation

The process is outlined by the following steps.

For any y (treat it as a deterministic variable):

1. Solve $g(x) = y$ for x. List *all* possible solutions $x_1, x_2, \ldots, x_n$.
2. Calculate $g'(x) = \frac{\mathrm{d}g(x)}{\mathrm{d}x}$.
3. Evaluate

$$f_y(y) = \begin{cases} 0, & \text{no solutions to } g(x) = y, \\ \sum_{i=1}^{n} \frac{f_x(x_i)}{|g'(x_i)|}, & \text{otherwise.} \end{cases} \tag{1.65}$$

The method is derived in [2, Sec. 2.6.2].

Example 1.5.2 (Method of transformation)**.** Consider

$$\underline{y} = g(\underline{x}) \tag{1.66}$$
$$= \underline{x}^2. \tag{1.67}$$

Carry the steps:

1. Solve for x in $y = g(x)$ for all possible cases of y:

   1.1. $y < 0$: No solutions.

1.2. $y = 0$: One solution $x_1 = 0$.

1.3. $y > 0$: Two solutions $x_1 = -\sqrt{y}$ and $x_2 = \sqrt{y}$.

2. Calculate the derivative of the function

$$g'(x) = 2x, \quad \forall x. \tag{1.68}$$

3. Calculate $f_y(y)$:

3.1. $y < 0$: $f_y(y) = 0$.

3.2. $y = 0$: $f_y(0) = 0$.

3.3. $y > 0$:

$$f_y(y) = \frac{f_x(-\sqrt{y})}{|-2\sqrt{y}|} + \frac{f_x(\sqrt{y})}{|2\sqrt{y}|} \tag{1.69}$$

$$= \frac{1}{2\sqrt{y}} \left[f_x(-\sqrt{y}) + f_x(\sqrt{y}) \right]. \tag{1.70}$$

△

1.5.3 Discrete transformation

The method will be motivated by two examples:

1. continuous to discrete transformation, and
2. discrete to discrete transformation.

Example 1.5.3 (Continuous to discrete transformation)**.** Let $\underline{x}$ be continuous with $f(x) = 0$, $x < 0$ and

$$\underline{y} = g(\underline{x}) \tag{1.71}$$

$$= \lfloor \underline{x} \rfloor \quad \text{(floor of } \underline{x}\text{)}. \tag{1.72}$$

Then the transformed variable $\underline{y}$ is discrete. I.e., $\mathcal{R}_y = \{0, 1, 2, \ldots\}$.

The PMF of $\underline{y}$ is then given by

$$p_y(y) = P\left(\underline{y} = y\right) \tag{1.73}$$

$$= P\left(\lfloor \underline{x} \rfloor = y\right) \tag{1.74}$$

$$= P\left(y \leq \underline{x} < y + 1\right) \tag{1.75}$$

$$= \int_y^{y+1} f_x(x) \mathrm{d}x, \quad y = 0, 1, 2, \ldots. \tag{1.76}$$

△

Example 1.5.4 (Discrete to discrete transformation)**.** Let $\underline{x}$ be discrete with $\mathcal{R}_x = \mathbb{Z}$, and

$$\underline{y} = g(\underline{x}) \tag{1.77}$$
$$= |\underline{x}| \,. \tag{1.78}$$

Then $\underline{y}$ is also discrete random variable with $\mathcal{R}_y = \{0, 1, 2, \ldots\}$.

The PMF of $\underline{y}$ can then be given by

$$p_y(y) = P\left(\underline{y} = y\right) \tag{1.79}$$
$$= P\left(|\underline{x}| = y\right) \tag{1.80}$$
$$= P\left(\underline{x} = \pm y\right) \tag{1.81}$$
$$= \begin{cases} p_x(0), & y = 0, \\ p_x(-y) + p_x(y), & y = 1, 2, \ldots . \end{cases} \tag{1.82}$$

△

1.6 Statistics of random variables

Sometimes, information about the random variables such as the mean is sought after. Information about a random variables are referred to as the *statistics* of it.

Definition 1.6.1 (*Expectation operator*)**.** The expectation operator $\mathbb{E}\left[\cdot\right]$ is a functional operator that operates on random variables. Specifically:

- For continuous random variables

$$\mathbb{E}\left[g(\underline{x})\right] = \int_{-\infty}^{\infty} g(x) f\left(x\right) \mathrm{d}x. \tag{1.83}$$

- For discrete random variables

$$\mathbb{E}\left[g(\underline{x})\right] = \sum_{-\infty}^{\infty} g(x) p\left(x\right). \tag{1.84}$$

Definition 1.6.2 (*Mean of a random variable*)**.** The mean of a random variable $\underline{x}$, μ_x, is defined as

$$\mu_x = \mathbb{E}\left[\underline{x}\right] \tag{1.85}$$

which

- for continuous random variable

$$\mu_x = \mathbb{E}[\underline{x}] \tag{1.86}$$
$$= \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x, \tag{1.87}$$

- while for discrete random variable

$$\mu_x = \mathbb{E}[\underline{x}] \tag{1.88}$$
$$= \sum_{-\infty}^{\infty} x p(x). \tag{1.89}$$

and the conditional mean of $\underline{x}$ given event M is

$$\mathbb{E}[\underline{x}|M] = \int_{-\infty}^{\infty} x f(x|M)\,\mathrm{d}x. \tag{1.90}$$

Properties of the mean

- Mean of $\underline{y} = g(\underline{x})$ is

$$\mathbb{E}[\underline{y}] = \mathbb{E}[g(\underline{x})] \tag{1.91}$$
$$= \int_{-\infty}^{\infty} g(x) f_x(x)\mathrm{d}x. \tag{1.92}$$

- Linear property:

$$\mathbb{E}\left[\sum_{i=1}^{n} \alpha_i g_i(\underline{x})\right] = \sum_{i=1}^{n} \alpha_i \mathbb{E}[g_i(\underline{x})]. \tag{1.93}$$

Note that in general,

$$\mathbb{E}[g(\underline{x})] \neq g(\mathbb{E}[\underline{x}]). \tag{1.94}$$

Definition 1.6.3 (*Variance and standard deviation*)**.** Let $\underline{x}$ be a random variable. Then the *variance* of $\underline{x}$, $\mathrm{Var}[\underline{x}]$, is defined as

$$\mathrm{Var}[\underline{x}] = \sigma_x^2 \tag{1.95}$$
$$= \mathbb{E}\left[(\underline{x} - \mathbb{E}[\underline{x}])^2\right] \tag{1.96}$$
$$= \mathbb{E}[\underline{x}^2] - \mathbb{E}[\underline{x}]^2. \tag{1.97}$$

The *standard deviation* of $\underline{x}$ is $\sigma_x = \sqrt{\mathrm{Var}[\underline{x}]}$.

A useful property of the variance is

$$\text{Var}\left[a\underline{x}+b\right]=a^2\,\text{Var}\left[\underline{x}\right]. \tag{1.98}$$

Theorem 1.6.1 (Markov's inequality)**.** Let $\underline{x}$ be a non-negative random variable. That is, $P\left(\underline{x}<0\right)=0$. For any $\epsilon>0$:

$$P\left(\underline{x}\geq\epsilon\right)\leq\frac{\mathbb{E}\left[\underline{x}\right]}{\epsilon}. \tag{1.99}$$

Theorem 1.6.2 (Chebyshev's inequality)**.** Let $\underline{x}$ be a random variable with mean μ and variance σ^2. Then, for any $\epsilon>0$:

$$P\left(|\underline{x}-\mu|\geq\epsilon\right)\leq\frac{\sigma^2}{\epsilon}. \tag{1.100}$$

Theorem 1.6.3 (Jensen's inequality)**.** Let $g(\cdot)$ be a convex function. Then,

$$\mathbb{E}\left[g(\underline{x})\right]\geq g(\mathbb{E}\left[\underline{x}\right]). \tag{1.101}$$

1.6.1 Moments

Definition 1.6.4 (*Moments*)**.** The different moments of a random variable are defined as follows.

- *n-th moment* of $\underline{x}$: $m_n=\mathbb{E}\left[\underline{x}^n\right]$.
- *n-th central moment* of $\underline{x}$: $\mu_n=\mathbb{E}\left[(\underline{x}-\mathbb{E}\left[\underline{x}\right])^n\right]$.
- *n-th absolute moment* of $\underline{x}$: $\mathbb{E}\left[|\underline{x}|^n\right]$.
- *n-th absolute central moment* of $\underline{x}$: $\mu_n=\mathbb{E}\left[|\underline{x}-\mathbb{E}\left[\underline{x}\right]|^n\right]$.

1.6.2 Characteristic functions

Definition 1.6.5 (*Characteristic function*)**.** The *characteristic function* of a random variable $\underline{x}$ is the Fourier transform of its PDF. Specifically,

$$\Phi_x(\omega)=\int_{-\infty}^{\infty}e^{\jmath\omega x}f_x(x)\mathrm{d}x, \tag{1.102}$$

where the inverse expression is given by

$$f_x(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-\jmath\omega x}\Phi_x(\omega)\mathrm{d}\omega. \tag{1.103}$$

Definition 1.6.6 (*Moment generating function*)**.** The *moment generating function* of $\underline{x}$ is the Laplace transform of its PDF. Specifically,

$$\bar{\Phi}_x(s) = \int_{-\infty}^{\infty} e^{Sx} f_x(x) \mathrm{d}x. \tag{1.104}$$

Theorem 1.6.4 (Moment theorem)**.**

$$\mathbb{E}\left[\underline{x}^n\right] = \left.\frac{\mathrm{d}^n}{\mathrm{d}s^n}\bar{\Phi}_x\right|_{s=0} = \left.\frac{1}{\jmath^n}\frac{\mathrm{d}^n}{\mathrm{d}\omega^n}\Phi_x(\omega)\right|_{\omega=0}. \tag{1.105}$$

Chapter 2

Multivariate Probability

2.1 Multivariate random variables and distributions

In this chapter, we are dealing with multiple random variables.

Definition 2.1.1 (*Multiple random variable*)**.** Let $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ be n random variables defined on the *same* probability space $(S, \mathcal{F}, P)$. Then, the mapping

$$s \in S \to [\underline{x}_1(s), \underline{x}_2(s), \ldots, \underline{x}_n(s)]^{\mathsf{T}} \in \mathbb{R}^n \tag{2.1}$$

defines an n-dimensional random variable (random vector or column matrix). Random vectors will be denoted by boldface underlined alphabet. Specifically,

$$\underline{\mathbf{x}} = \begin{bmatrix} \underline{x}_1 & \underline{x}_2 & \ldots & \underline{x}_n \end{bmatrix}^{\mathsf{T}}, \tag{2.2}$$

where $\underline{x}_i, i = 1, \ldots, n$ are random variables.

2.1.1 Joint cumulative distribution function (JCDF)

Definition 2.1.2 (*Joint cumalitive distribution function (JCDF)*)**.** Let $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ be n random variables defined on the same probability space $(S, \mathcal{F}, P)$. Then, the *joint cumulative distribution function (JCDF)* of these random variables are

$$\begin{aligned} F_{x_1,x_2,\ldots,x_n}(x_1, x_2, \ldots, x_n) &= P\left(\underline{x}_1 \leq x_1, \underline{x}_2 \leq x_2, \ldots, \underline{x}_n \leq x_n\right) && (2.3)\\ &= F_{\mathbf{x}}(\mathbf{x}) && (2.4)\\ &= F\left(\mathbf{x}\right) && (2.5)\\ &= P\left(\underline{\mathbf{x}} \leq \mathbf{x}\right). && (2.6) \end{aligned}$$

Properties of JCDFs

1. $F\left(\mathbf{x}\right)$ is non-decreasing in each of its arguments.

2. $F(\mathbf{x})$ is right-continuous in each of its arguments.
3. For a fixed i, $\lim_{x_i \to -\infty} F(\mathbf{x}) = 0$.
4. When $\underline{x}_i \to \infty$ for *all* i, then $F(\mathbf{x}) \to 1$.

Definition 2.1.3 (*Marginal CDF*)**.** Let $\underline{\mathbf{x}} = \begin{bmatrix} \underline{\mathbf{x}}_1^\mathsf{T} & \underline{\mathbf{x}}_2^\mathsf{T} \end{bmatrix}^\mathsf{T} \in \mathbb{R}^n$ with joint CDF $F(\mathbf{x})$. Say we want to know the joint CDF of a subset of these random variables $\underline{\mathbf{x}}_1$. That is, we want to *marginalize out* the random variables $\underline{\mathbf{x}}_2$. The joint CDF of $\underline{\mathbf{x}}_1$ is computed by taking the limit

$$F(\mathbf{x}_1) = \lim_{\mathbf{x}_2 \to \infty} F(\mathbf{x}). \tag{2.7}$$

Here are the definitions:

- The process is called *marginalization*.
- The retained random variables $\underline{\mathbf{x}}_1$ are called *marginal random variables*.
- The joint CDF of the marginal random variables is called the *marginal CDF*.
- The discarded random variables $\underline{\mathbf{x}}_2$ is said to be *marginalized out*.

Example 2.1.1 (Marginalizing a CDF)**.** Given the joint CDF $F(x_1, x_2, x_3)$, then the joint CDF $F(x_1, x_3)$ is given by

$$F(x_1, x_3) = \lim_{x_2 \to \infty} F(x_1, x_2, x_3). \tag{2.8}$$

△

2.1.2 Joint probability mass function (JPMF)

Definition 2.1.4 (*Joint probability mass function (JPMF)*)**.** This is the analog to joint CDF for discrete random variables. The *joint probability mass function (JPMF)* of the discrete random variables $\underline{x}_1, \ldots, \underline{x}_n$ is

$$\begin{aligned} p_{x_1,\ldots,x_n}(x_1, \ldots, x_n) &= P(\underline{x}_1 = x_1, \ldots, \underline{x}_n = x_n) && (2.9)\\ &= p_{\mathbf{x}}(\mathbf{x}) && (2.10)\\ &= P(\underline{\mathbf{x}} = \mathbf{x}). && (2.11) \end{aligned}$$

Note that the subscripts may be omitted. That is, $p_{\mathbf{x}}(\mathbf{x}) = p(\mathbf{x})$.

Properties of JPMFs

1. **Total probability**: $0 \leq p(\mathbf{x}) \leq 1$.

2. **Normalization Property**:

$$\sum_{\mathbf{x}\in\mathcal{R}_{\mathbf{x}}} p(\mathbf{x}) = \sum_{x_1\in\mathcal{R}_{x_1}} \cdots \sum_{x_n\in\mathcal{R}_{x_n}} p(x_1,\ldots,x_n) \tag{2.12}$$

$$= 1. \tag{2.13}$$

3. Probability that $\underline{\mathbf{x}}$ is in a region $\mathcal{D}$ (i.e., $\underline{\mathbf{x}} \in \mathcal{D}$) is given by

$$P(\underline{\mathbf{x}} \in \mathcal{D}) = \sum_{\underline{\mathbf{x}}\in\mathcal{D}} p(\mathbf{x}). \tag{2.14}$$

4. **Marginalization property**: Let $\underline{\mathbf{x}} = \begin{bmatrix}\underline{\mathbf{x}}_1^\mathsf{T} & \underline{\mathbf{x}}_2^\mathsf{T}\end{bmatrix}^\mathsf{T}$ be a discrete random variable with joint PMF $p(\mathbf{x})$. Then the joint PMF of $\underline{\mathbf{x}}_1$ is given by marginalization. Specifically,

$$p(\mathbf{x}_1) = \sum_{\mathbf{x}_2\in\mathcal{R}_{\mathbf{x}_2}} p(\mathbf{x}). \tag{2.15}$$

2.1.3 Joint probability density function (JPDF)

Definition 2.1.5 (*Joint probability density function (JPDF)*)**.** The *joint PDF* of random variables $\underline{\mathbf{x}} = \begin{bmatrix}\underline{x}_1 & \cdots & \underline{x}_n\end{bmatrix}^\mathsf{T}$, $f(\mathbf{x})$, is given by

$$f(\mathbf{x}) = f(x_1,\ldots,x_n) \tag{2.16}$$

$$= \frac{\partial^n F(x_1,\ldots,x_n)}{\partial x_1\cdots\partial x_n}. \tag{2.17}$$

Often, subscripts will be dropped (i.e., $f_{\mathbf{x}}(\mathbf{x}) = f(\mathbf{x})$).

Properties of JPDFs

1. $f(\mathbf{x}) \geq 0$.

2. **Normalization**:

$$\int_{\mathbb{R}^n} f(\mathbf{x}) = \int_{\mathbb{R}^n} f(x_1,\ldots,x_n)\,\mathrm{d}x_1\cdots\mathrm{d}x_n \tag{2.18}$$

$$= 1. \tag{2.19}$$

3. Probability that $\underline{\mathbf{x}}$ is in a region $\mathcal{D}$:

$$P(\underline{\mathbf{x}} \in \mathcal{D}) = \int_{\mathcal{D}} f(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{2.20}$$

4. **Marginalization property**: Let $\underline{\mathbf{x}} = \begin{bmatrix}\underline{\mathbf{x}}_1^\mathsf{T} & \underline{\mathbf{x}}_n^\mathsf{T}\end{bmatrix}^\mathsf{T}$ be a continuous random variable with joint PDF

$f(\mathbf{x})$. Then the joint PDF of $\underline{\mathbf{x}}_1$ is

$$f(\mathbf{x}_1) = \int_{-\infty}^{\infty} f(\mathbf{x})\, \mathrm{d}\mathbf{x}. \tag{2.21}$$

2.1.4 Independence

Definition 2.1.6 (*Mutual independence*)**.** The random variables $\underline{x}_1, \ldots, \underline{x}_n$ are called *mutually independent* if the events $\{\underline{x}_1 \leq x_1\}, \ldots, \{\underline{x}_n \leq x_n\}$ are independent for any $x_1, \ldots, x_n \in \mathbb{R}^n$.

It follows that

$$F(x_1, \ldots, x_n) = F(x_1) \cdots F(x_n), \tag{2.22}$$
$$f(x_1, \ldots, x_n) = f(x_1) \cdots f(x_n), \tag{2.23}$$
$$p(x_1, \ldots, x_n) = p(x_1) \cdots p(x_n). \tag{2.24}$$

Remark 2.1.1. Here some remarks about mutual independence.

- Any subset of the mutually independent random variables $\{\underline{x}_1, \ldots, \underline{x}_n\}$ is a set of mutually independent random variables.
- If $\{\underline{x}_1, \ldots, \underline{x}_n\}$ are independent *in pairs*, they are not necessarily independent.
- The random variables $\underline{y}_1 = g_1(\underline{x}_1), \ldots, \underline{y}_n = g_n(\underline{x}_n)$ are independent if $\{\underline{x}_1, \ldots, \underline{x}_n\}$ are independent. Note that the functions $g_1, \ldots, g_n$ need not be the same.
- Care must be taken when talking about independence. For instance, "independence" of a group of random variables may signify pairwise independence which does not imply mutual independence.

Definition 2.1.7 (*Group independence*)**.** A group of random variables $\mathcal{G}_{\mathbf{x}} = \{\underline{x}_1, \ldots, \underline{x}_n\}$ is *independent* of a group of random variables $\mathcal{G}_{\mathbf{y}} = \left\{\underline{y}_1, \ldots, \underline{y}_m\right\}$ if

$$f(\mathbf{x}, \mathbf{y}) = f(\mathbf{x})\, f(\mathbf{y}). \tag{2.25}$$

Definition 2.1.8 (*Independent and identically distributed (i.i.d.) random variables*)**.** The random variables $\underline{x}_1, \ldots, \underline{x}_n$ are called *independent and identically distributed (i.i.d.)* if

1. they are independent, and
2. $F_{x_1}(x) = F_{x_2}(x) = \cdots = F_{x_n}(x)$, or equivalently

$$f_{x_1}(x) = \cdots = f_{x_n}(x), \quad \text{or} \tag{2.26}$$
$$p_{x_1}(x) = \cdots = p_{x_n}(x). \tag{2.27}$$

2.1.4.1 Complex random variables

The statistics (PDF, CDF, etc.) of the n complex random variables

$$\begin{aligned} \underline{z}_1 &= \underline{x}_1 + \jmath \underline{y}_1 \\ &\vdots \\ \underline{z}_n &= \underline{x}_n + \jmath \underline{y}_n \end{aligned} \tag{2.28}$$

are determined by the joint PDF $f(x_1, \ldots, x_n, y_1, \ldots, y_n)$ of the $2n$ real random variables $\mathbf{x}, \mathbf{y}$.

Definition 2.1.9 (*Mutually independent complex random variables*)**.** The complex random variables $\underline{z}_1, \ldots, \underline{z}_n$ as defined in (2.28) are *mutually independent* if

$$f(\mathbf{x}, \mathbf{y}) = f(x_1, y_1) \cdots f(x_n, x_n). \tag{2.29}$$

2.2 Transformed random variables

Consider n random variables $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ with joint PDF $f(x_1, \ldots, x_n)$. Further, given a function $g : \mathbb{R}^n \to \mathbb{R}^k$

$$g(\mathbf{x}) = \begin{bmatrix} g_1(\mathbf{x}) \\ \vdots \\ g_k(\mathbf{x}) \end{bmatrix}. \tag{2.30}$$

Then there are k new random variables $\underline{\mathbf{y}} = \begin{bmatrix} \underline{y}_1 & \cdots & \underline{y}_k \end{bmatrix}^{\mathsf{T}}$. Specifically,

$$\underline{\mathbf{y}} = g(\underline{\mathbf{x}}). \tag{2.31}$$

Basic concepts

If $\underline{\mathbf{x}}$ is continuous, then

$$\begin{aligned} F(\mathbf{y}) &= F(y_1, \ldots, y_k) && (2.32) \\ &= P\left(\underline{y}_1 \leq y_1, \ldots, \underline{y}_k \leq y_k\right) && (2.33) \\ &= P\left(g_1(\underline{\mathbf{x}}) \leq y_1, \ldots, g_k(\underline{\mathbf{x}}) \leq y_k\right) && (2.34) \\ &= \int_{\mathcal{D}} f(\mathbf{x})\, \mathrm{d}\mathbf{x}, && (2.35) \end{aligned}$$

where

$$\begin{aligned} \mathcal{D} &= \{\mathbf{x} \in \mathbb{R}^n \text{ s.t. } g_1(\mathbf{x}) \leq y_1, \ldots, g_k(\mathbf{x}) \leq y_k\} && (2.36) \\ &= \{x_1, \ldots, x_n \text{ s.t. } g_1(\cdot) \leq y_1, \ldots, g_k(\cdot) \leq y_k\}. && (2.37) \end{aligned}$$

That is, $\mathcal{D}$ depends on $\mathbf{y}$.

If $\underline{\mathbf{x}}$ is discrete, then

$$p(\mathbf{y}) = P(y_1, \ldots, y_k) \tag{2.38}$$
$$= \sum_{\mathbf{x} \in \mathcal{D}} P(x_1, \ldots, x_n) \tag{2.39}$$

where

$$\mathcal{D} = \{\mathbf{x} \in \mathbb{R}^n \text{ s.t. } g_1(\mathbf{x}) = y_1, \ldots, g_k(\mathbf{x}) = y_k\} \tag{2.40}$$
$$= \{x_1, \ldots, x_n \text{ s.t. } g_1(\cdot) = y_1, \ldots, g_k(\cdot) = y_k\}. \tag{2.41}$$

Again, $\mathcal{D}$ depends on $\mathbf{y}$.

Transformed variables distribution

Given the distribution of $\underline{\mathbf{x}}$ (CDF, PDF, or PMF) and the function $g : \mathbb{R}^k \to \mathbb{R}^n$, then what is the distribution of $\underline{\mathbf{y}} = g(\underline{\mathbf{x}}) \in \mathbb{R}^k$?

Assumption: $k \leq n$. That is, the mapping g is surjective.

Just as is the case for a single transformed random variable (Section 1.5), there are different methods to find the distribution of a transformed random variable (i.e., $\underline{y} = g(\underline{x})$) depending on the types of variables.

1. **Method of distributions** works for *any* type of random variables (continuous or discrete). However, it is the most exhaustive method.
2. **Method of transformations** works for *continuous* random variables. Specifically, both $\underline{\mathbf{x}}$ and $\underline{\mathbf{y}}$ are continuous.

2.2.1 Method of distribution

Given the function $g : \mathbb{R}^n \to \mathbb{R}^k$, where $k \leq n$, then carry the following.

1. For all $\mathbf{y} \in \mathbb{R}^k$, find $\mathbf{x} \in D_y \subseteq \mathbb{R}^n$ such that

$$g(\mathbf{x}) \leq \mathbf{y}. \tag{2.42}$$

   Note that D_y is a function of $\mathbf{y}$.

2. The CDF of $\underline{\mathbf{y}}$ is then

$$F(\mathbf{y}) = P\left(\underline{\mathbf{y}} \leq \mathbf{y}\right) \tag{2.43}$$
$$= P(\underline{\mathbf{x}} \in D_y) \tag{2.44}$$
$$= \int_{D_y} f(\mathbf{x})\, \mathrm{d}\mathbf{x}. \tag{2.45}$$

3. The PDF of $\underline{\mathbf{y}}$ is then

$$f(\mathbf{y}) = \frac{\mathrm{d}F(\mathbf{y})}{\mathrm{d}\mathbf{y}}. \tag{2.46}$$

Example 2.2.1. Consider the transformed random variable

$$\underline{y} = g(\underline{x}_1, \underline{x}_2) \tag{2.47}$$
$$= \underline{x}_1 + \underline{x}_2. \tag{2.48}$$

To find the PDF of $\underline{y}$, the steps outlined above will be follows.

1. Find D_y:

$$D_y = \left\{(x_1, x_2) \in \mathbb{R}^2 : x_2 \leq y - x_1\right\}. \tag{2.49}$$

2. CDF of $\underline{y}$:

$$F(y) = \int\int_{D_y} f(x_1, x_2)\,\mathrm{d}x_1\mathrm{d}x_2 \tag{2.50}$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{y-x_1} f(x_1, x_2)\,\mathrm{d}x_2\mathrm{d}x_1. \tag{2.51}$$

3. PDF of $\underline{y}$ is

$$f(y) = \frac{\mathrm{d}F(y)}{\mathrm{d}y} \tag{2.52}$$
$$= \int_{-\infty}^{\infty} \frac{\mathrm{d}}{\mathrm{d}y}\int_{-\infty}^{y-x_1} f(x_1, x_2)\,\mathrm{d}x_2\mathrm{d}x_1 \tag{2.53}$$
$$= \int_{-\infty}^{\infty} f(x_1, y - x_1)\,\mathrm{d}x_1. \tag{2.54}$$

△

Theorem 2.2.1 (Transformed variable PDF of linear combination of random variables)**.** Let $\underline{x}_1$ and $\underline{x}_2$ be jointly distributed random variables with joint PDF $f(x_1, x_2)$. The PDF of $\underline{y} = \underline{x}_1 + \underline{x}_2$ is given by

$$f(y) = \int_{-\infty}^{\infty} f(\lambda, y - \lambda)\,\mathrm{d}\lambda. \tag{2.55}$$

Theorem 2.2.2 (Transformed variable PDF of linear combination of independent random variables)**.** Let $\underline{x}_1$ and $\underline{x}_2$ be *independent* random variables with marginal PDFs $f_1(x_1)$ and $f_2(x_2)$, respectively. Then,

the PDF of $\underline{y} = \underline{x}_1 + \underline{x}_2$ is given by

$$f(y) = \int_{-\infty}^{\infty} f_1(\lambda) f_2(y - \lambda) \mathrm{d}\lambda \tag{2.56}$$

$$= f_1(y) * f_2(y), \tag{2.57}$$

where $*$ denotes the convolution operator.

2.2.2 Method of transformation

Note that the method is derived in [2, Sec. 2.6.2].

2.2.2.1 Approach 1

Given the transformed random variable

$$\underline{\mathbf{y}} = g(\underline{\mathbf{x}}), \tag{2.58}$$

where $\underline{\mathbf{y}} \in \mathbb{R}^k$ (i.e., $g : \mathbb{R}^n \to \mathbb{R}^k$) and given the PDF of the random variable $\underline{\mathbf{x}}$. For now, **assume** $k = n$. The distribution of $\underline{\mathbf{y}}$ is then found as follows.

1. For all $\mathbf{y}$, solve the system of equations

$$\mathbf{y} = g(\mathbf{x}) \tag{2.59}$$

for $\mathbf{x}$. For a given $\mathbf{y}$, the system has either

1.1. m solutions

$$\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}, \tag{2.60}$$

1.2. or no solutions.

Note that the system will not have infinite solutions since it is assumed that $k = n$.

2. Evaluate the Jacobian of g

$$J(\mathbf{x}) = \frac{\mathrm{d}g(\mathbf{x})}{\mathrm{d}\mathbf{x}}. \tag{2.61}$$

3. Then, the PDF of $\underline{\mathbf{y}}$ is

$$f(\mathbf{y}) = \sum_{i=1}^{m} \frac{f\left(\mathbf{x}^{(i)}\right)}{\left|\det\left(J(\mathbf{x}^{(i)})\right)\right|}, \tag{2.62}$$

where $\det(\cdot)$ is the determinant of a square matrix and $|\cdot|$ is the absolute value. **If for y the system**

has no solution, then

$$f(\mathbf{y}) = \mathbf{0}. \tag{2.63}$$

2.2.2.2 Approach 2

Given the transformed random variable

$$\underline{\mathbf{y}} = g(\underline{\mathbf{x}}), \tag{2.64}$$

where $\underline{\mathbf{y}} \in \mathbb{R}^k$ (i.e., $g : \mathbb{R}^n \to \mathbb{R}^k$) and given the PDF of the random variable $\mathbf{x}$. For now, **assume** $k = n$. The distribution of $\underline{\mathbf{y}}$ is then found as follows.

1. Identical to in Step 1 of approach 1 (Section 2.2.2.1): for all $\mathbf{y}$, solve the system of equations

   $$\mathbf{y} = g(\mathbf{x}) \tag{2.65}$$

   for $\mathbf{x}$. For a given $\mathbf{y}$, the system has either

   1.1. m solutions

   $$\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}, \tag{2.66}$$

   1.2. or no solutions.

   Note that the system will not have infinite solutions since it is assumed that $k = n$.

2. Evaluate the Jacobian

   $$\tilde{J}_i(\mathbf{y}) = \frac{\mathrm{d}\mathbf{x}^{(i)}}{\mathrm{d}\mathbf{y}} \tag{2.67}$$

   $$= \begin{bmatrix} \frac{\partial x_1^{(i)}}{\partial y_1} & \cdots & \frac{\partial x_1^{(i)}}{\partial y_n} \\ \vdots & & \vdots \\ \frac{\partial x_n^{(i)}}{\partial y_1} & \cdots & \frac{\partial x_n^{(i)}}{\partial y_n} \end{bmatrix} \tag{2.68}$$

   for each solution $i = 1, \ldots, m$.

3. Then, the PDF of $\underline{\mathbf{y}}$ is

   $$f(y) = \sum_{i=1}^{m} \left| \det\left(\tilde{J}_i(\mathbf{x}^{(i)}) \right) \right| f\left(\mathbf{x}^{(i)} \right), \tag{2.69}$$

   where $\det(\cdot)$ is the determinant of a square matrix and $|\cdot|$ is the absolute value. **If for y the system has no solution**, then

   $$f(\mathbf{y}) = \mathbf{0}. \tag{2.70}$$

Example 2.2.2. Consider the transformed random variable $\underline{\mathbf{y}}$ given by

$$\underline{\mathbf{y}} = g(\underline{\mathbf{x}}) \tag{2.71}$$

$$= \begin{bmatrix} \underline{x}_1^2 \\ \underline{x}_1 + \underline{x}_2 \end{bmatrix}. \tag{2.72}$$

1. For all $\mathbf{y}$ solve

$$\mathbf{y} = g(\mathbf{x}) \tag{2.73}$$

$$= \begin{bmatrix} \underline{x}_1^2 \\ \underline{x}_1 + \underline{x}_2 \end{bmatrix} \tag{2.74}$$

for $\mathbf{x}$.

Case I $(y_1 < 0)$: No solutions.

Case II $(y_1 = 0)$: One solution

$$\mathbf{x} = \begin{bmatrix} 0 \\ y_2 \end{bmatrix}. \tag{2.75}$$

Case III $(y_1 > 0)$: two solutions:

$$\mathbf{x}^{(1)} = \begin{bmatrix} \sqrt{y_1} \\ y_2 - \sqrt{y_1} \end{bmatrix} \tag{2.76}$$

$$\mathbf{x}^{(2)} = \begin{bmatrix} -\sqrt{y_1} \\ y_2 + \sqrt{y_1} \end{bmatrix}. \tag{2.77}$$

2. For the ith solution (using approach 1):

$$\det\left(J_i(\mathbf{x})\right) = \det\left(\frac{\mathrm{d}g(\mathbf{x})}{\mathrm{d}\mathbf{x}}\right) \tag{2.78}$$

$$= \det\left(\begin{bmatrix} \frac{\partial x_1^{(i)}}{\partial y_1} & \frac{\partial x_1^{(i)}}{\partial y_2} \\ \frac{\partial x_2^{(i)}}{\partial y_1} & \frac{\partial x_2^{(i)}}{\partial y_2} \end{bmatrix}\right) \tag{2.79}$$

$$= \frac{\partial x_1^{(i)}}{\partial y_1}\frac{\partial x_2^{(i)}}{\partial y_2} - \frac{\partial x_2^{(i)}}{\partial y_1}\frac{\partial x_1^{(i)}}{\partial y_2}. \tag{2.80}$$

Case I $(y_1 < 0)$: no solutions.

Case II $(y_1 = 0)$: one solution:

$$\det\left(J(\mathbf{x})\right) = 0. \tag{2.81}$$

Case III ($y_1 > 0$): two solutions:

$$J_1(\mathbf{x}) = \frac{1}{2\sqrt{y_1}} \tag{2.82}$$

$$J_2(\mathbf{x}) = -\frac{1}{2\sqrt{y_1}}. \tag{2.83}$$

3. The JPDF of $\underline{\mathbf{y}}$ is then

$$f_{\mathbf{y}}(\mathbf{y}) = \sum_i f_{\mathbf{x}}(\mathbf{x}) \left|\det\left(J_i\right)\right| \tag{2.84}$$

or $f_{\mathbf{y}}(\mathbf{y}) = \mathbf{0}$ if there are no solutions.

$$f_{\mathbf{y}}(\mathbf{y}) = f_{\mathbf{x}^{(1)}}(\mathbf{x}) \left|\det J_1\right| + f_{\mathbf{x}^{(2)}}(\mathbf{x}) \left|\det J_2\right| \tag{2.85}$$

$$= f_{\mathbf{x}}(\sqrt{y_1}, y_2 - \sqrt{y_1}) \left|\frac{1}{2\sqrt{y_1}}\right| + f_{\mathbf{x}}(-\sqrt{y_1}, y_2 + \sqrt{y_1}) \left|-\frac{1}{2\sqrt{y_1}}\right|. \tag{2.86}$$

△

So far, in this method, we assumed that $k = n$. But if $k < n$, then follow the following steps.

1. Introduce $n - k$ "dummy" variables[a]:

$$\underline{y}_{k+1} = g_{k+1}(\underline{\mathbf{x}}) \tag{2.87}$$

$$\vdots \tag{2.88}$$

$$\underline{y}_n = g_n(\underline{\mathbf{x}}). \tag{2.89}$$

2. Evaluate $f\left(y_1, \ldots, y_k, \ldots, y_n\right)$ using the methods outlined before (now we have $k = n$).
3. Marginalize out $\underline{y}_{k+1}, \ldots, \underline{y}_n$ to obtain $f\left(\underline{y}_1, \ldots, \underline{y}_k\right)$.

[a] I believe it's better if the introduced mapping is bijective.

2.2.3 Linear transformation

If $\mathbf{g}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ (i.e., $\mathbf{g}$ is linear), then the mean and the covariance are given by

$$\mathbb{E}\left[\underline{\mathbf{y}}\right] = \mathbf{A}\bar{\mathbf{x}}, \tag{2.90}$$

$$\mathrm{Cov}\left[\underline{\mathbf{y}}\right] = \mathbb{E}\left[\left(\underline{\mathbf{y}} - \bar{\mathbf{y}}\right)\left(\underline{\mathbf{y}} - \bar{\mathbf{y}}\right)^{\mathsf{T}}\right] \tag{2.91}$$

$$= \mathbb{E}\left[\mathbf{A}\left(\underline{\mathbf{x}} - \bar{\mathbf{x}}\right)\left(\mathbf{A}(\underline{\mathbf{x}} - \bar{\mathbf{x}})\right)^{\mathsf{T}}\right] \tag{2.92}$$

$$= \mathbf{A}\mathbb{E}\left[\left(\underline{\mathbf{x}} - \bar{\mathbf{x}}\right)\left(\underline{\mathbf{x}} - \bar{\mathbf{x}}\right)^{\mathsf{T}}\right]\mathbf{A}^{\mathsf{T}} \tag{2.93}$$

$$= \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^{\mathsf{T}}. \tag{2.94}$$

Transform	Acronym	Benefits	Drawbacks
First order Taylor expansion	TT1	Simple and fast to compute.	Inaccurate for highly nonlinear functions.
Second order Taylor expansion	TT2	Better accuracy than TT1.	Requires computation of Hessian of the function.
Monte-Carlo simulation	MCT	Very accurate.	Requires many samples, thus computationally inefficient.
Unscented transform	UCT	Accuracy close to TT2	Doesn't require second order information (Hessian).

Table 2.1: Summary of transforms that approximate the covariances of propagated random variables.

Note that the relation is an exact relation, not an approximation.

2.2.4 Nonlinear transformation (covariance propagation)

For the nonlinear case, getting the exact transformed covariance is tedious and cumbersome. Therefore, tools that approximate the covariance are used that trade off between accuracy and computational efficiency. Most of the information from this section is obtained from [3]. Table 2.1 summarizes the results.

2.2.4.1 First order Taylor expansion

For a nonlinear transformation, a simple approximation is linearization. Thus,

$$\underline{\mathbf{y}} \approx \mathbf{g}(\bar{\mathbf{x}}) + \mathbf{g}'(\bar{\mathbf{x}})\delta\underline{\mathbf{x}}. \tag{2.95}$$

The random variable becomes $\delta\underline{\mathbf{x}} \sim \mathcal{N}(\underline{\mathbf{x}} - \bar{\mathbf{x}}, \mathbf{\Sigma}_x)$. The covariance of $\underline{\mathbf{y}}$ is then given by

$$\begin{aligned}
\mathbb{E}\left[\underline{\mathbf{y}}\right] &= \mathbf{g}(\bar{\mathbf{x}}), && (2.96)\\
\mathrm{Cov}\left[\underline{\mathbf{y}}\right] &= \mathbb{E}\left[\left(\underline{\mathbf{y}} - \bar{\mathbf{y}}\right)\left(\underline{\mathbf{y}} - \bar{\mathbf{y}}\right)^{\mathsf{T}}\right] && (2.97)\\
&= \mathbb{E}\left[\left(\underline{\mathbf{y}} - \mathbf{g}(\bar{\mathbf{x}})\right)\left(\underline{\mathbf{y}} - \mathbf{g}(\bar{\mathbf{x}})\right)^{\mathsf{T}}\right] && (2.98)\\
&= \mathbb{E}\left[\mathbf{g}'(\bar{\mathbf{x}})\delta\underline{\mathbf{x}}\delta\underline{\mathbf{x}}^{\mathsf{T}}\mathbf{g}'(\bar{\mathbf{x}})^{\mathsf{T}}\right] && (2.99)\\
&= \mathbf{g}'(\bar{\mathbf{x}})\mathbf{\Sigma}_x\mathbf{g}'(\bar{\mathbf{x}})^{\mathsf{T}}. && (2.100)
\end{aligned}$$

2.2.4.2 Monte-Carlo simulation

Monte-Carlo simulation consists of simply sampling a *large* number (say in the order of $N = 10^4$) of the domain random variable $\underline{\mathbf{x}}$ and passing it through the nonlinear function $\mathbf{g}$ to get a large number of samples of $\underline{\mathbf{y}}$, $\mathbf{y}^{(i)}, i = 1, \ldots, N$. The mean and covariance are then given by

$$\bar{\mathbf{y}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{y}^{(i)}, \tag{2.101}$$

$$\mathrm{Cov}\left[\bar{\mathbf{y}}\right] \approx \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{y}^{(i)} - \bar{\mathbf{y}}\right)\left(\mathbf{y}^{(i)} - \bar{\mathbf{y}}\right)^{\mathsf{T}}, \tag{2.102}$$

where the $\frac{1}{N-1}$ term is to ensure that the estimator is *unbiased*.

2.2.4.3 Unscented transform

Let n_x be the length of the domain random variable (i.e., $\underline{\mathbf{x}} \in \mathbb{R}^{n_x}$). Further, let $\mathbf{u}_i \in \mathbb{R}^{n_x}$ be the left singular vector of the singular value σ_i of the domain covariance matrix $\mathbf{\Sigma}_x$. That is,

$$\mathbf{\Sigma}_x = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^{\mathsf{T}} \tag{2.103}$$

$$= \sum_{i=1}^{n_x} \sigma_i^2 \mathbf{u}_i \mathbf{u}_i^{\mathsf{T}}, \tag{2.104}$$

where $\mathbf{\Sigma}$ is a diagonal matrix of the singular values.

Let $\mathbf{x}^{(i)}$ be the i-th realization of $\underline{\mathbf{x}}$. For this version of the unscented transform (UT), the number of realizations will be $2n_x + 1$. Specifically, set

$$\mathbf{x}^{(0)} = \bar{\mathbf{x}}, \tag{2.105}$$

$$\mathbf{x}^{(\pm i)} = \bar{\mathbf{x}} \pm \sqrt{n_x + \lambda}\sigma_i \mathbf{u}_i, \qquad i = 1, \ldots, n_x, \tag{2.106}$$

where λ is a user-defined parameter to be discussed and $\bar{\mathbf{x}}$ is the mean of $\underline{\mathbf{x}}$ (in estimation problems, this is usually the predicted value). Further, let

$$\omega^{(0)} = \frac{\lambda}{n_x + \lambda}, \tag{2.107}$$

$$\omega^{(\pm i)} = \frac{1}{2(n_x + \lambda)}, \qquad i = 1, \ldots, n_x. \tag{2.108}$$

Next, set

$$\mathbf{y}^{(\pm i)} = \mathbf{g}(\mathbf{x}^{(\pm i)}), i = 0, \ldots, n_x. \tag{2.109}$$

Then, the stats of the propagated random variable $\underline{\mathbf{y}}$ are

$$\bar{\mathbf{y}} = \sum_{i=-n_x}^{n_x} \omega^{(i)} \mathbf{y}^{(i)}, \tag{2.110}$$

$$\mathbf{\Sigma}_y = \left(1 - \alpha^2 + \beta\right) \left(\mathbf{y}^{(0)} - \bar{\mathbf{y}}\right) \left(\mathbf{y}^{(0)} - \bar{\mathbf{y}}\right)^{\mathsf{T}} \tag{2.111}$$

$$+ \sum_{i=-n_x}^{n_x} \omega^{(i)} \left(\mathbf{y}^{(i)} - \bar{\mathbf{y}}\right) \left(\mathbf{y}^{(i)} - \bar{\mathbf{y}}\right)^{\mathsf{T}}, \tag{2.112}$$

where α and β are user-defined parameters to be discussed. Further, $\omega^{(0)} + (1 - \alpha^2 + \beta)$ is often denoted $\omega_c^{(0)}$. The UT discussed above is the 'mod' version of the UT. There is another version denoted 'std' obtained by removing the first term in (2.111). I have no idea about the differences between the 'std' and 'mod' versions.

User-defined parameters

The authors of [3] discuss the values to set the user-defined parameters.

- α controls the spread of the sigma points and is suggested to be approximately 10^{-3}.

- β compensates for the distribution (not really sure what that means), and should be chosen as $\beta = 2$ when $\underline{\mathbf{x}}$ is Gaussian.
- λ is defined by $\lambda = \alpha^2 (n_x + \kappa) - n_x$, where κ is usually zero.
- $\omega^{(0)} = 1 - \frac{n_x}{3}$ for the 'std' version of the UT when $\underline{\mathbf{x}}$ is Gaussian.

2.2.4.4 Spherical cubature method

Given a function a random variable $\underline{\mathbf{x}} \in \mathbb{R}^{n_x}$ and a nonlinear function $\mathbf{g} : \mathbb{R}^{n_x} \to \mathbb{R}^{n_y}$, then the spherical cubature method[1] approximates the mean and the covariance of the transformed random variable $\underline{\mathbf{y}} = \mathbf{g}(\underline{\mathbf{x}})$. The method is summarized in Algorithm 1. More details are available in [4, Algorithm 6.8]

Algorithm 1 Nonlinear transformation: The spherical cubature

1: **Input**: The distribution on the random variable $\underline{\mathbf{x}} \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ where $\mathbf{x} \in \mathbb{R}^{n_x}$, and the nonlinear function $\mathbf{g} : \mathbb{R}^{n_x} \to \mathbb{R}^{n_y}$.

2: Generate the unit sigma points using

$$\boldsymbol{\xi} = \sqrt{n_x} \begin{bmatrix} \mathbf{1} & -\mathbf{1} \end{bmatrix} \in \mathbb{R}^{n_x \times 2 * n_x}. \tag{2.113}$$

3: Generate the sigma points

$$\mathcal{X}^{(i)} = \boldsymbol{\mu}_x + \sqrt{\boldsymbol{\Sigma}_x}\boldsymbol{\xi}^{(i)}, \quad , i = 1, \ldots, 2n_x, \tag{2.114}$$

where $\boldsymbol{\xi}^{(i)} \in \mathbb{R}^{n_x}$ is the i-th column of $\boldsymbol{\xi}$, and $\boldsymbol{\Sigma}_x = \sqrt{\boldsymbol{\Sigma}_x}\sqrt{\boldsymbol{\Sigma}_x}^{\mathsf{T}}$ is obtained using Cholesky decomposition.

4: Pass the sigma-points through the nonlinearity

$$\mathcal{Y}^{(i)} = \mathbf{g}(\mathcal{X}^{(i)}), \quad i = 1, \ldots, 2n_x. \tag{2.115}$$

5: Compute

$$\boldsymbol{\mu}_y = \frac{1}{2n_x} \sum_{i=1}^{2n_x} \mathcal{Y}^{(i)}, \tag{2.116}$$

$$\boldsymbol{\Sigma}_y = \frac{1}{2n_x} \sum_{i=1}^{2n_x} \left(\mathcal{Y}^{(i)} - \boldsymbol{\mu}_y\right) \left(\mathcal{Y}^{(i)} - \boldsymbol{\mu}_y\right)^{\mathsf{T}}. \tag{2.117}$$

2.3 Statistics of random variables

2.3.1 Order statistics

Definition 2.3.1 (*Order statistics*)**.** Let $\underline{x}_1, \ldots, \underline{x}_n$ be i.i.d. random variables with marginal CDF and

[1] It's a type of a sigma-point transformation.

PDF $F(\cdot)$ and $f(\cdot)$, respectively. Consider the transformation

$$\underline{x}_{(1)} = \text{ smallest value of } (\underline{x}_1, \ldots, \underline{x}_n) \tag{2.118}$$

$$\underline{x}_{(2)} = \text{ 2nd smallest value of } (\underline{x}_1, \ldots, \underline{x}_n) \tag{2.119}$$

$$\vdots \tag{2.120}$$

$$\underline{x}_{(n)} = n\text{th smallest (i.e., largest) value of} (\underline{x}_1, \ldots, \underline{x}_n), \tag{2.121}$$

then $\underline{x}_{(1)}, \ldots, \underline{x}_{(n)}$ are the *order statistics* of $\underline{x}_1, \ldots, \underline{x}_n$. Furthermore, $\underline{x}_{(k)}$ is the *kth order statistic* of $\underline{x}_1, \ldots, \underline{x}_n$.

Theorem 2.3.1 (CDF of order statistics)**.** The CDF of the kth order statistic of $\underline{x}_1, \ldots, \underline{x}_n$ is

$$F_{(k)}(x) = \sum_{i=k}^{n} \binom{n}{i} F(x)^i \left(1 - F(x)\right)^{n-i}. \tag{2.122}$$

2.3.2 Mean

Theorem 2.3.2 (Mean of a function)**.** The *mean* of a function $g : \mathbb{R}^n \to \mathbb{R}$ of the n real random variables $\underline{x}_1, \ldots, \underline{x}_n$ is given by

$$\mathbb{E}\left[g(\mathbf{x})\right] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\mathbf{x}) f(\mathbf{x}) \, \mathrm{d}\mathbf{x}, \tag{2.123}$$

where $\mathbb{E}[\cdot]$ is the multidimensional expectation operator (Definition 1.6.1).

Theorem 2.3.3 (Mean of a function of complex random variables)**.** Mean of a function $g : \mathbb{R}^n \to \mathbb{R}$ of the n complex random variables $\underline{z}_i = \underline{x}_i + \jmath \underline{y}_i, i = 1, \ldots, n$ is given by

$$\mathbb{E}\left[g(\underline{z}_1, \ldots, \underline{z}_n)\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1 + \jmath y_1, \ldots, x_n + \jmath y_n) f(x_1, \ldots, x_n, y_1, \ldots, y_n) \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}, \tag{2.124}$$

where $\mathbb{E}[\cdot]$ is the multidimensional expectation operator (Definition 1.6.1).

Definition 2.3.2 (*Mean of a random vector*)**.** The *mean* of a random vector $\underline{\mathbf{x}} = \begin{bmatrix} \underline{x}_1 & \cdots & \underline{x}_n \end{bmatrix}^{\mathsf{T}}$ is defined as

$$\mathbb{E}\left[\underline{\mathbf{x}}\right] = \begin{bmatrix} \mathbb{E}\left[\underline{x}_1\right] & \cdots & \mathbb{E}\left[\underline{x}_n\right] \end{bmatrix}^{\mathsf{T}}. \tag{2.125}$$

Note that the mean of a random vector is also a vector of the same length. That is, $\mathbb{E}\left[\underline{\mathbf{x}}\right] \in \mathbb{R}^n$.

Definition 2.3.3 (*Mean of a random matrix*)**.** Let

$$\underline{\mathbf{X}} = \begin{bmatrix} \underline{\mathbf{x}}_1^{\mathsf{T}} \\ \vdots \\ \underline{\mathbf{x}}_m^{\mathsf{T}} \end{bmatrix}_{m \times n}, \tag{2.126}$$

be a random matrix, where $\underline{\mathbf{x}}_i \in \mathbb{R}^n, i = 1, \ldots, m$. Then, the mean of $\underline{\mathbf{X}}$ is given by

$$\mathbb{E}\left[\underline{\mathbf{X}}\right] = \begin{bmatrix} \mathbb{E}\left[\underline{\mathbf{x}}_1\right]^{\mathsf{T}} \\ \vdots \\ \mathbb{E}\left[\underline{\mathbf{x}}_m\right]^{\mathsf{T}} \end{bmatrix}. \tag{2.127}$$

Properties of the expectation operator
The expectation operator is linear. That is

$$\mathbb{E}\left[\alpha_1 g_1(\underline{\mathbf{x}}) + \ldots + \alpha_m g_m(\underline{\mathbf{x}})\right] = \alpha_1 \mathbb{E}\left[g_1(\underline{\mathbf{x}})\right] + \ldots + \alpha_m \mathbb{E}\left[g_m(\underline{\mathbf{x}})\right] \tag{2.128}$$

for any $\underline{\mathbf{x}}$ (real or complex) and any *deterministic* $\alpha_1, \ldots, \alpha_m$ (real or complex). As a result,

$$\mathbb{E}\left[\mathbf{A}\underline{\mathbf{x}}\right] = \mathbf{A}\mathbb{E}\left[\underline{\mathbf{x}}\right], \tag{2.129}$$

for any deterministic matrix $\mathbf{A}$.

2.3.3 Covariance and correlation

Definition 2.3.4 (*Correlation matrix*)**.** The *correlation matrix* of the (complex) random variable $\underline{\mathbf{x}} = \begin{bmatrix} \underline{x}_1 & \cdots & \underline{x}_n \end{bmatrix}^{\mathsf{T}}$ is defined as

$$\mathbf{R} = \mathbb{E}\left[\underline{\mathbf{x}}\underline{\mathbf{x}}^{\mathsf{H}}\right], \tag{2.130}$$

where $(\cdot)^{\mathsf{H}}$ is the Hermitian[a] of a vector. The (i, j)th element of $\mathbf{R}$ is the correlation of $\underline{x}_i$ and $\underline{x}_j$.

[a] Hermitian: $\mathbf{x}^{\mathsf{H}} = (\mathbf{x}^*)^{\mathsf{T}} = (\mathbf{x}^{\mathsf{T}})^*$.

Definition 2.3.5 (*Covariance matrix*)**.** The *covariance matrix* of the random vector $\underline{\mathbf{x}} = \begin{bmatrix} \underline{x}_1 & \cdots & \underline{x}_n \end{bmatrix}^{\mathsf{T}}$ is

$$\mathbf{C} = \mathbb{E}\left[\left(\underline{\mathbf{x}} - \mathbb{E}\left[\underline{\mathbf{x}}\right]\right)\left(\underline{\mathbf{x}} - \mathbb{E}\left[\underline{\mathbf{x}}\right]\right)^{\mathsf{H}}\right]. \tag{2.131}$$

The (i, j)th element of $\mathbf{C}$ is the covariance of $\underline{x}_i$ and $\underline{x}_j$. That is,

$$C_{ij} = \mathbb{E}\left[\left(\underline{x}_i - \mathbb{E}\left[\underline{x}_i\right]\right)\left(\underline{x}_j - \mathbb{E}\left[\underline{x}_j\right]\right)^*\right]. \tag{2.132}$$

The covariance matrix $\mathbf{C}$ of $\underline{\mathbf{x}}$ is equal to the correlation matrix of the "centered" random vector $\underline{\mathbf{x}} - \mathbb{E}[\underline{\mathbf{x}}]$.

Definition 2.3.6 (*Mutually orthogonal random variables*)**.** The random variables $\underline{x}_1, \ldots, \underline{x}_n$ are called *mutually orthogonal* if the correlation matrix $\mathbf{R}$ is diagonal. That is, $R_{ij} = 0, i \neq j$.

Definition 2.3.7 (*Mutually uncorrelated random variables*)**.** The random variables $\underline{x}_1, \ldots, \underline{x}_n$ are called *mutually uncorrelated* if the covariance matrix $\mathbf{C}$ is diagonal. That is, $C_{ij} = 0, i \neq j$.

Notes: If the random variables $\underline{x}_1, \ldots, \underline{x}_n$ are independent then they are uncorrelated. The converse is not true in general.

Definition 2.3.8 (*Uncorrelated random vectors*)**.** Two random vectors $\underline{\mathbf{x}}$ and $\underline{\mathbf{y}}$ are called *uncorrelated* if

$$\mathbb{E}\left[\left(\underline{\mathbf{x}} - \mathbb{E}[\underline{\mathbf{x}}]\right)\left(\underline{\mathbf{y}} - \mathbb{E}\left[\underline{\mathbf{y}}\right]\right)^{\mathsf{H}}\right] = \mathbf{0}. \tag{2.133}$$

Or equivalently

$$\mathbb{E}\left[\underline{\mathbf{x}}\underline{\mathbf{y}}^{\mathsf{H}}\right] = \mathbb{E}[\underline{\mathbf{x}}]\,\mathbb{E}\left[\underline{\mathbf{y}}\right]^{\mathsf{H}}. \tag{2.134}$$

Definition 2.3.9 (*Orthogonal random vectors*)**.** Two random vectors $\underline{\mathbf{x}}$ and $\underline{\mathbf{y}}$ are called orthogonal if

$$\mathbb{E}\left[\underline{\mathbf{x}}\underline{\mathbf{y}}^{\mathsf{H}}\right] = \mathbf{0}. \tag{2.135}$$

Properties

1. Relationship between the correlation and covariance matrix

$$\begin{aligned}\mathbf{C} &= \mathbf{R} - \mathbb{E}[\underline{\mathbf{x}}]\,\mathbb{E}[\underline{\mathbf{x}}]^{\mathsf{H}} && (2.136)\\ &= \mathbb{E}\left[\underline{\mathbf{x}}\underline{\mathbf{x}}^{\mathsf{H}}\right] - \mathbb{E}[\underline{\mathbf{x}}]\,\mathbb{E}[\underline{\mathbf{x}}]^{\mathsf{H}}. && (2.137)\end{aligned}$$

2. The correlation and covariance matrix are positive semidefinite matrices.

   Proof. For any vector $\mathbf{a}$,

$$\begin{aligned}\mathbf{a}^{\mathsf{H}}\mathbf{R}\mathbf{a} &= \mathbf{a}^{\mathsf{H}}\mathbb{E}\left[\underline{\mathbf{x}}\underline{\mathbf{x}}^{\mathsf{H}}\right]\mathbf{a} && (2.138)\\ &= \mathbb{E}\left[\left(\mathbf{a}^{\mathsf{H}}\underline{\mathbf{x}}\right)\left(\mathbf{a}^{\mathsf{H}}\underline{\mathbf{x}}\right)^{\mathsf{H}}\right] && (2.139)\\ &= \mathbb{E}\left[\left|\mathbf{a}^{\mathsf{H}}\underline{\mathbf{x}}\right|^2\right] && (2.140)\\ &\geq 0. && (2.141)\end{aligned}$$

□

3. The eigenvalues of the correlation and covariance matrix are real and non-negative.

Theorem 2.3.4. If $\mathbf{R}$ is the correlation matrix of the n-dimensional random vector $\underline{\mathbf{x}}$. Then,

$$\mathbb{E}\left[\underline{\mathbf{x}}^{\mathsf{H}}\mathbf{R}^{-1}\underline{\mathbf{x}}\right] = n. \tag{2.142}$$

Proof. Since $\underline{\mathbf{x}}^{\mathsf{H}}\mathbf{R}^{-1}\underline{\mathbf{x}}$ is a scalar, we have

$$\begin{aligned}
\mathbb{E}\left[\underline{\mathbf{x}}^{\mathsf{H}}\mathbf{R}^{-1}\underline{\mathbf{x}}\right] &= \mathbb{E}\left[\operatorname{tr}\left(\underline{\mathbf{x}}^{\mathsf{H}}\mathbf{R}^{-1}\underline{\mathbf{x}}\right)\right] && (2.143)\\
&= \mathbb{E}\left[\operatorname{tr}\left(\underline{\mathbf{x}}\underline{\mathbf{x}}^{\mathsf{H}}\mathbf{R}^{-1}\right)\right] && (2.144)\\
&= \operatorname{tr}\mathbb{E}\left[\underline{\mathbf{x}}\underline{\mathbf{x}}^{\mathsf{H}}\mathbf{R}^{-1}\right] && (2.145)\\
&= \operatorname{tr}\left(\mathbb{E}\left[\underline{\mathbf{x}}\underline{\mathbf{x}}^{\mathsf{H}}\right]\mathbf{R}^{-1}\right) && (2.146)\\
&= \operatorname{tr}\left(\mathbf{R}\mathbf{R}^{-1}\right) && (2.147)\\
&= \operatorname{tr}\mathbf{1} && (2.148)\\
&= n. && (2.149)
\end{aligned}$$

□

2.4 Moments and characteristic functions

Definition 2.4.1 (*Moments of random vectors*)**.** Consider the random variables $\underline{x}_1, \ldots, \underline{x}_N$ and $n_i = 0, 1, \ldots$.

- The $(n_1, n_2, \ldots, n_N)$th *moment* of $\underline{x}_1, \ldots, \underline{x}_N$ is

$$\mathbb{E}\left[\underline{x}_1^{n_1} \cdots \underline{x}_N^{n_N}\right]. \tag{2.150}$$

  Example ($N = 3$): (2,3,1)th moment is $\mathbb{E}\left[\underline{x}_1^2 \cdot \underline{x}_2^3 \cdot \underline{x}_3^1\right]$.

- The $(n_1, \ldots, n_N)$th *central moment* of $\underline{x}_1, \ldots, \underline{x}_N$ is

$$\mathbb{E}\left[(\underline{x}_1 - \mathbb{E}\left[\underline{x}_1\right])^{n_1} \cdots (\underline{x}_N - \mathbb{E}\left[\underline{x}_N\right])^{n_N}\right]. \tag{2.151}$$

- The $(n_1, \ldots, n_N)$th *absolute moment* of $\underline{x}_1, \ldots, \underline{x}_N$ is

$$\mathbb{E}\left[|\underline{x}_1|^{n_1} \cdots |\underline{x}_N|^{n_N}\right]. \tag{2.152}$$

Definition 2.4.2 (*Characteristic function*)**.** The *characteristic function* of a random vector $\underline{\mathbf{x}}$ is defined as

$$\Phi_{\mathbf{x}}(\Omega) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{\jmath(\omega_1 x_1 + \ldots + \omega_N x_N)} f\left(x_1, \ldots, x_N\right) \mathrm{d}x_1 \cdots \mathrm{d}x_N \tag{2.153}$$

$$= \mathbb{E}\left[e^{\jmath \Omega^{\mathsf{T}} \underline{\mathbf{x}}}\right] \tag{2.154}$$

where $\Omega = \begin{bmatrix} \omega_1 & \cdots & \omega_N \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^N$ and $\underline{\mathbf{x}} = \begin{bmatrix} \underline{x}_1 & \cdots & \underline{x}_N \end{bmatrix}^{\mathsf{T}}$. The inverse relation is given by

$$f\left(x_1, \ldots, x_N\right) = \frac{1}{(2\pi)^N} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \Phi_{\mathbf{x}}(\Omega) e^{-\jmath(\omega_1 x_1 + \ldots + \omega_N x_N)} \mathrm{d}\omega_1 \cdots \mathrm{d}\omega_N. \tag{2.155}$$

Note that it looks similar to the Fourier transform of the PDF but it's not exactly the same because of the sign in the exponential.

Definition 2.4.3 (*Moment generating function*)**.** The *moment generating function* of a random vector $mbfrvx$ is defined as

$$\mathrm{Var}\left[\Phi\right]_{\mathbf{x}}(\mathbf{s}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{s_1 x_1 + \ldots + s_N x_N} f\left(x_1, \ldots, x_N\right) \mathrm{d}x_1 \cdots \mathrm{d}x_N \tag{2.156}$$

$$= \mathbb{E}\left[e^{\mathbf{s}^{\mathsf{T}} \underline{\mathbf{x}}}\right] \tag{2.157}$$

where

$$\mathbf{s} = \begin{bmatrix} s_1 & \cdots & s_N \end{bmatrix}^{\mathsf{T}}. \tag{2.158}$$

Theorem 2.4.1 (Characteristic functions)**.** The $(n_1, \ldots, n_N)$th *moment* of $\underline{\mathbf{x}}$ is given by

$$m_{n_1,\ldots,n_N} = \mathbb{E}\left[\underline{x}_1^{n_1 \cdots} \underline{x}_N^{n_N}\right] \tag{2.159}$$

$$= \frac{\partial^{n_1 + \cdots + n_N}}{\partial s_1^{n_1} \cdots \partial s_N^{n_N}} \bar{\Phi}_{\mathbf{x}}(\mathbf{s}) \bigg|_{s_1 = \cdots = s_N = 0} \tag{2.160}$$

$$= \frac{1}{\jmath^{n_1 + \cdots + n_N}} \frac{\partial^{n_1 + \cdots + n_N}}{\partial \omega_1^{n_1} \cdots \partial \omega_N^{n_N}} \bar{\Phi}_{\mathbf{x}}(\Omega) \bigg|_{\omega_1 = \cdots = \omega_N = 0}. \tag{2.161}$$

Theorem 2.4.2 (PDF of independent random variables)**.** Let $\underline{x}_1, \ldots, \underline{x}_n$ be independent random variables with marginal PDFs $f_i(x), i = 1, \ldots, n$, respectively. The PDF of $\underline{y} = \underline{x}_1 + \cdots + \underline{x}_n$ is

$$f_y(y) = f_1(y) * \cdots * f_n(y), \tag{2.162}$$

where $*$ denotes the convolution.

2.5 Conditional distributions

Similar to the single random variable definition of conditional distributions (Definition 1.4.6), the following are extensions for such definitions to the multivariate case.

Definition 2.5.1 (*CDF of continuous RV conditioned on discrete RV*)**.** Let $\underline{x}$ and $\underline{y}$ be two random variables with $\underline{x}$ *discrete*. The *conditional CDF of* $\underline{y}$ *given* $\underline{x} = x$ is defined as

$$F\left(y|\ \underline{x} = x\right) = P\left(\underline{y} \leq y \middle|\ \underline{x} = x\right). \tag{2.163}$$

Remark 2.5.1. Some remarks about CDFs.

- Conditions: $p\left(x\right) = P\left(\underline{x} = x\right) \neq 0$.
- Notation: $F\left(y|\ \underline{x} = x\right) \to F_{y|x}(y|x)$, or $F\left(y|x\right)$.

Definition 2.5.2 (*PMF of discrete RV conditioned a discrete RV*)**.** Let $\underline{x}$ and $\underline{y}$ be two joint discrete random variables. The *conditional PMF of* $\underline{y}$ *given* $\underline{x} = x$ is defined as

$$p\left(y|\ \underline{x} = x\right) = P\left(\underline{y} = y \middle|\ \underline{x} = x\right) \tag{2.164}$$

$$= \frac{p_{xy}(x, y)}{p_x(x)}. \tag{2.165}$$

Notation: $p\left(y|\ \underline{x} = x\right) \to p_{y|x}(y|x)$, or $p\left(y|x\right)$.

Definition 2.5.3 (*PDF of continuous RV conditioned on continuous RV*)**.** Let $\underline{x}$ and $\underline{y}$ be two jointly continuous random variables. The *conditional PDF of* $\underline{y}$ *given* $\underline{x} = x$ is defined as

$$f\left(y|\ \underline{x} = x\right) = \frac{f_{xy}(x, y)}{f_x(x)}. \tag{2.166}$$

Remark 2.5.2. Some remarks about PDFs.

- Note: if $\underline{x}$ is continuous, then $P\left(\underline{x} = x\right) = 0$. Therefore, it is not possible to use a PDF definition similar to that of a PMF.
- Notation: $f\left(y|\ \underline{x} = x\right) \to f_{y|x}(y|x)$, or $f\left(y|x\right)$.

Definition 2.5.4 (*CDF of continous RV conditioned on continuous RV*)**.** Let $\underline{x}$, $\underline{y}$ be two jointly continuous random variables. The *conditional CDF of* $\underline{y}$ *given* $\underline{x} = x$ is defined as

$$F\left(y|\ \underline{x} = x\right) = \int_{-\infty}^{y} f_{y|x}(\lambda|x)\mathrm{d}\lambda. \tag{2.167}$$

- Notation: $F\left(y|\underline{x} = x\right) \to F_{y|x}(y|x)$, or $F\left(y|x\right)$.

Remark 2.5.3. Some remarks about CDFs.

1. Note that the conditional CDF is derived from the conditional PDF, rather than the other way around (deriving PDF from CDF as was done in the unconditional case).
2. For a given x, $f(y|x)$ is a valid PDF. That is,
   2.1. $f(y|x) \geq 0$
   2.2. $\int_{-\infty}^{\infty} f(y|x)\,\mathrm{d}y = 1$.
3. $f(x|y) = \frac{f(x,y)}{f(y)} = \frac{f(y|x)f(x)}{f(y)}$.
4. If $\underline{x}$ and $\underline{y}$ are independent, then $f(x,y) = f(x)\,f(y)$. Therefore,

$$f(y|x) = f(y), \tag{2.168}$$
$$f(x|y) = f(x). \tag{2.169}$$

Definition 2.5.5 (*Multivaraite conditional PDF*)**.** The *multivariate conditional PDF* of $\underline{x}_n, \ldots, \underline{x}_{k+1}$ given $\underline{k} = x_k, \ldots, \underline{x}_1 = x_1$ is

$$f(x_N, \ldots, x_{k+1}|\, x_k, \ldots, x_1) = \frac{f(x_n, \ldots, x_{k+1}, x_k, \ldots, x_1)}{f(x_k, \ldots, x_1)} \tag{2.170}$$

Definition 2.5.6 (*Multivariate conditional CDF*)**.** The *multivariate conditional CDF* of $\underline{x}_n, \ldots, \underline{x}_{k+1}$ given $\underline{k} = x_k, \ldots, \underline{x}_1 = x_1$ is

$$F(x_N, \ldots, x_{k+1}|\, x_k, \ldots, x_1) = \tag{2.171}$$
$$\int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_{k+1}} f_{x_n,\ldots,x_{k+1}|x_k,\ldots,x_1}(\lambda_n, \ldots, \lambda_{k+1}|\, x_k, \ldots, x_1)\mathrm{d}\lambda_n \cdots \mathrm{d}\lambda_{k+1}. \tag{2.172}$$

2.5.1 Properties of conditional probability

1. **Total probability**:

$$f(y) = \int_{-\infty}^{\infty} f(y|x)\,f(x)\,\mathrm{d}x. \tag{2.173}$$

2. **Bayes Theorem**:

$$f(x|y) = \frac{f(y|x)\,f(x)}{\int_{-\infty}^{\infty} f(y|x)\,f(x)\,\mathrm{d}x}. \tag{2.174}$$

3. **Chain rule**:

$$f(x_1, \ldots, x_n) = f(x_n|x_{n-1}, \ldots, x_1) \cdots f(x_3|x_2, x_1)\,f(x_2|x_1)\,f(x_1). \tag{2.175}$$

4. **Marginalization**: Removing any number of variables on the left of the conditional line can be done by integrating the PDF with respect to them. For example,

$$f\left(\mathbf{x}_1 \middle|\, \mathbf{x}_3\right) = \int_{\infty}^{\infty} f\left(\mathbf{x}_1, \mathbf{x}_2 \middle|\, \mathbf{x}_3\right) \mathrm{d}\mathbf{x}_2. \tag{2.176}$$

5. **Chapman-Kolmogoroff equation**: Removing any number of variables to the right of the conditional line can be done by multiplying ty their conditional density given the remaining variables on the right, and then integrating the product. That is,

$$f\left(\mathbf{x}_1 \middle|\, \mathbf{x}_3\right) = \int_{-\infty}^{\infty} f\left(\mathbf{x}_1 \middle|\, \mathbf{x}_2, \mathbf{x}_3\right) f\left(\mathbf{x}_2 \middle|\, \mathbf{x}_3\right) \mathrm{d}\mathbf{x}_2. \tag{2.177}$$

2.5.2 Conditional covariance

These equations are obtained from [5, Eq. (2.53b)].

Given a the random variable

$$\underline{\mathbf{x}} = \begin{bmatrix} \underline{\mathbf{x}}_1 \\ \underline{\mathbf{x}}_2 \end{bmatrix} \sim \mathcal{N}\left(\bar{\mathbf{x}}, \boldsymbol{\Sigma}\right), \tag{2.178}$$

where

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{\mathbf{x}}_1 \\ \bar{\mathbf{x}}_2 \end{bmatrix}, \tag{2.179}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^{\mathsf{T}} & \boldsymbol{\Sigma}_{22} \end{bmatrix}. \tag{2.180}$$

Then, what is the distribution of $\underline{\mathbf{x}}_1$ given $\underline{\mathbf{x}}_2 = \hat{\mathbf{x}}_2$?

The Schur's compliment can be used to compute this value. Specifically,

$$\mathbb{E}\left[\underline{\mathbf{x}}_1 | \underline{\mathbf{x}}_2 = \hat{\mathbf{x}}_2\right] = \bar{\mathbf{x}}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\hat{\mathbf{x}}_2 - \bar{\mathbf{x}}_2\right), \tag{2.181}$$

$$\mathrm{Cov}\left[\underline{\mathbf{x}}_1 | \underline{\mathbf{x}}_2 = \hat{\mathbf{x}}_2\right] = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^{\mathsf{T}}. \tag{2.182}$$

2.6 Conditional expectation and variance

Definition 2.6.1 (*Conditional mean*)**.** Given n random variables $\underline{x}_1, \ldots, \underline{x}_n$, the *conditional mean* of the random variable $h(\underline{x}_1, \ldots, \underline{x}_k)$ conditioned on $\underline{x}_{k+1}, \ldots, \underline{x}_n$ is

$$\mathbb{E}\left[h(\underline{x}_1, \ldots, \underline{x}_k) \middle|\, x_{k+1}, \ldots, x_n\right] = \tag{2.183}$$

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x_1, \ldots, x_k) f\left(x_1, \ldots, x_k \middle|\, x_{k+1}, \ldots, x_n\right) \mathrm{d}x_1 \cdots \mathrm{d}x_k. \tag{2.184}$$

Remark 2.6.1. Some remarks about conditional means.

- Special case:

$$\mathbb{E}\left[\underline{x}_1 \middle| x_2, \ldots, x_n\right] = \int_{-\infty}^{\infty} x_1 f\left(x_1 \middle| x_2, \ldots, x_n\right) \mathrm{d}x_1. \tag{2.185}$$

- If $\underline{x}_1, \ldots, \underline{x}_n$ are independent of $\underline{x}_{k+1}, \ldots, \underline{x}_n$, then

$$\mathbb{E}\left[h(\underline{x}_1, \ldots, \underline{x}_k) \middle| x_{k+1}, \ldots, x_n\right] = \mathbb{E}\left[h(\underline{x}_1, \ldots, \underline{x}_k)\right]. \tag{2.186}$$

Theorem 2.6.1 (Nested expectation)**.** Let $\underline{x}$, $\underline{y}$ be two jointly distributed RVs. Further, let

$$g(y) = \mathbb{E}\left[\underline{x} \middle| \underline{y} = y\right] \tag{2.187}$$

$$= \int_{-\infty}^{\infty} x f\left(x|y\right) \mathrm{d}x. \tag{2.188}$$

Then

$$\mathbb{E}\left[g(\underline{y})\right] = \mathbb{E}\left[\underline{x}\right]. \tag{2.189}$$

In other words,

$$\mathbb{E}_y\left[\mathbb{E}_{x|y}\left[\underline{x} \middle| \underline{y}\right]\right] \to \mathbb{E}\left[\mathbb{E}\left[\underline{x} \middle| \underline{y}\right]\right] \tag{2.190}$$

$$= \mathbb{E}\left[\underline{x}\right]. \tag{2.191}$$

Proof.

$$\mathbb{E}\left[g(\underline{y})\right] = \int_{-\infty}^{\infty} g(y) f\left(y\right) \mathrm{d}y \tag{2.192}$$

$$= \int_{-\infty}^{\infty} \mathbb{E}\left[\underline{x} \middle| y\right] f\left(y\right) \mathrm{d}y \tag{2.193}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f\left(x|y\right) f\left(y\right) \mathrm{d}x \mathrm{d}y \tag{2.194}$$

$$= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f\left(x, y\right) \mathrm{d}y \mathrm{d}x \tag{2.195}$$

$$= \int_{-\infty}^{\infty} x f\left(x\right) \mathrm{d}x \tag{2.196}$$

$$= \mathbb{E}\left[\underline{x}\right]. \tag{2.197}$$

□

Theorem 2.6.2 (Multivariate nested expectation)**.** Let $\underline{x}_1, \ldots, \underline{x}_n$ be n random variables. Also, let

$$g(x_2, \ldots, x_n) = \mathbb{E}\left[\underline{x}_1 \middle| x_2, \ldots, x_n\right]. \tag{2.198}$$

Then,

$$\mathbb{E}\left[g(\underline{x}_2, \ldots, \underline{x}_n)\right] = \mathbb{E}\left[\underline{x}_1\right]. \tag{2.199}$$

Notation:

$$\mathbb{E}_{x_2,\ldots,x_n}\left[\mathbb{E}_{x_1|x_2,\ldots,x_n}\left[\underline{x}_1 \middle| \underline{x}_2, \ldots, \underline{x}_n\right]\right] \to \mathbb{E}\left[\mathbb{E}\left[\underline{x}_1 \middle| \underline{x}_2, \ldots, \underline{x}_n\right]\right]. \tag{2.200}$$

Theorem 2.6.3 (Multivariate nested expectation: general case)**.** Let $\underline{x}_1, \ldots, \underline{x}_n$ be n random variables and $h : \mathbb{R}^k \to \mathbb{R}$. Also, let

$$g(x_{k+1}, \ldots, x_n) = \mathbb{E}\left[h(\underline{x}_1, \ldots, \underline{x}_k) \middle| x_{k+1}, \ldots, x_n\right]. \tag{2.201}$$

Then,

$$\begin{aligned}\mathbb{E}\left[g(\underline{x}_{k+1}, \ldots, \underline{x}_n)\right] &= \mathbb{E}\left[\mathbb{E}\left[h(\underline{x}_1, \ldots, \underline{x}_k)\middle| \underline{x}_{k+1}, \ldots, \underline{x}_n\right]\right] &(2.202)\\ &= \mathbb{E}\left[h(\underline{x}_1, \ldots, \underline{x}_k)\right]. &(2.203)\end{aligned}$$

Definition 2.6.2 (*Conditional variance*)**.** Let $\underline{x}$, $\underline{y}$ be two random variables. Then the *conditional variance of* $\underline{x}$ *given* $\underline{y} = y$ is

$$\mathrm{Var}\left[\underline{x}\middle| \underline{y} = y\right] = \mathbb{E}\left[\left(\underline{x} - \mathbb{E}\left[\underline{x}\middle| \underline{y} = y\right]\right)^2 \middle| \underline{y} = y\right]. \tag{2.204}$$

2.6.1 Properties of conditional variance

1. $\mathrm{Var}\left[\underline{x}\middle| \underline{y} = y\right] = \mathbb{E}\left[\underline{x}^2\middle| \underline{y} = y\right] - \mathbb{E}\left[\underline{x}\middle| \underline{y} = y\right]^2$.
2. $\mathrm{Var}\left[\underline{x}\middle| \underline{y} = y\right]$ is a function of y.
3. $g(\underline{y}) = \mathrm{Var}\left[\underline{x}\middle| \underline{y}\right]$ is a random variable.

Theorem 2.6.4 (Nested conditional variance)**.** Let $\underline{x}$, $\underline{y}$ be two random variables. Then

$$\mathrm{Var}\left[\underline{x}\right] = \mathbb{E}\left[\mathrm{Var}\left[\underline{x}\middle| \underline{y}\right]\right] + \mathrm{Var}\left[\mathbb{E}\left[\underline{x}\middle| \underline{y}\right]\right]. \tag{2.205}$$

Proof.

$$\mathbb{E}\left[\mathrm{Var}\left[\underline{x}\middle|\,\underline{y}\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\underline{x}^2\middle|\,\underline{y}\right] - \mathbb{E}\left[\underline{x}\middle|\,\underline{y}\right]^2\right] \tag{2.206}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\underline{x}^2\middle|\,\underline{y}\right]\right] - \mathbb{E}\left[\mathbb{E}\left[\underline{x}\middle|\,\underline{y}\right]^2\right] \tag{2.207}$$

$$= \mathbb{E}\left[\underline{x}^2\right] - \mathbb{E}\left[\mathbb{E}\left[\underline{x}\middle|\,\underline{y}\right]^2\right]. \tag{2.208}$$

Further,

$$\mathrm{Var}\left[\mathbb{E}\left[\underline{x}\middle|\,\underline{y}\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\underline{x}\middle|\,\underline{y}\right]^2\right] - \mathbb{E}\left[\mathbb{E}\left[\underline{x}\middle|\,\underline{y}\right]^2\right] \tag{2.209}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\underline{x}\middle|\,\underline{y}\right]^2\right] - \mathbb{E}\left[\underline{x}\right]^2. \tag{2.210}$$

Adding these two results proves the theorem. □

2.7 Passing measurements through a function

Theorem 2.7.1 (Passing measurements through a function)**.** Let $\underline{\mathbf{y}} \in \mathbb{R}^m$ be a (vector) random variable that depends on $\underline{\mathbf{x}} \in \mathbb{R}^n$ through

$$\underline{\mathbf{y}} = \mathbf{g}\left(\underline{\mathbf{x}}\right). \tag{2.211}$$

Furthermore, let $\underline{\mathbf{z}} \in \mathbb{R}^p$ be a transformed random variable given by

$$\underline{\mathbf{z}} = \mathbf{f}\left(\underline{\mathbf{y}}\right). \tag{2.212}$$

Then

$$f\left(\underline{\mathbf{x}}\middle|\,\mathbf{y},\mathbf{z}\right) = f\left(\underline{\mathbf{x}}\middle|\,\mathbf{y}\right). \tag{2.213}$$

Furthermore, if $\mathbf{g}$ is invertible and $\mathbf{f}$ is invertible, then

$$f\left(\underline{\mathbf{x}}\middle|\,\mathbf{y},\mathbf{z}\right) = f\left(\underline{\mathbf{x}}\middle|\,\mathbf{z}\right). \tag{2.214}$$

Note that if $\mathbf{g}$ is invertible but $\mathbf{f}$ is non-invertible, then (2.214) does not hold in general!

Proof. 1. The first part.

$$f(\mathbf{x}|\,\mathbf{y},\mathbf{z}) = \frac{f(\mathbf{y},\mathbf{z}|\,\mathbf{x})\,f(\mathbf{x})}{f(\mathbf{y},\mathbf{z})} \tag{2.215}$$

$$= \frac{f(\mathbf{z}|\,\mathbf{y},\mathbf{x})\,f(\mathbf{y}|\,\mathbf{x})\,f(\mathbf{x})}{f(\mathbf{z}|\,\mathbf{y})\,f(\mathbf{y})} \tag{2.216}$$

$$= \frac{\cancel{f(\mathbf{z}|\,\mathbf{y})}f(\mathbf{y}|\,\mathbf{x})\,f(\mathbf{x})}{\cancel{f(\mathbf{z}|\,\mathbf{y})}f(\mathbf{y})} \tag{2.217}$$

$$= \frac{f(\mathbf{y}|\,\mathbf{x})\,f(\mathbf{x})}{f(\mathbf{y})} \tag{2.218}$$

$$= f(\mathbf{x}|\,\mathbf{y}). \tag{2.219}$$

2. The second part. Let's expand $f(\mathbf{x}|\,\mathbf{y},\mathbf{z})$.

$$f(\mathbf{x}|\,\mathbf{y},\mathbf{z}) = \frac{f(\mathbf{y},\mathbf{z}|\,\mathbf{x})\,f(\mathbf{x})}{f(\mathbf{y},\mathbf{z})} \tag{2.220}$$

$$= \frac{f(\mathbf{y}|\,\mathbf{z},\mathbf{x})\,f(\mathbf{z}|\,\mathbf{x})\,f(\mathbf{x})}{f(\mathbf{y}|\,\mathbf{z})\,f(\mathbf{z})} \tag{2.221}$$

$$= \frac{f(\mathbf{y}|\,\mathbf{z},\mathbf{x})}{f(\mathbf{y}|\mathbf{z})}f(\mathbf{x}|\,\mathbf{z}). \tag{2.222}$$

Thus, (2.214) satisfied if and only if

$$f(\mathbf{y}|\,\mathbf{z},\mathbf{x}) = f(\mathbf{y}|\mathbf{z}). \tag{2.223}$$

That is, if $\mathbf{y}$ is independent of $\mathbf{x}$ *given* $\mathbf{z}$. More rigorously, if there exists $\mathbf{y}_1 \neq \mathbf{y}_2$ such that

$$\left.\begin{aligned}\mathbf{g}(\mathbf{x}_1) &= \mathbf{y}_1\\ \mathbf{g}(\mathbf{x}_2) &= \mathbf{y}_2\end{aligned}\right\} \implies \mathbf{x}_1 \neq \mathbf{x}_2, \tag{2.224}$$

and

$$\mathbf{z} = \mathbf{f}(\mathbf{y}_1) = \mathbf{f}(\mathbf{y}_2). \tag{2.225}$$

Then knowing $\mathbf{z}$ without knowing $\mathbf{x}$ will result in a non-singleton sample space S for the random variable $\underline{\mathbf{y}}$. That is, a set with more than one element which implies that $\underline{\mathbf{y}}$ cannot be determined uniquely. In such case, the sample space is the set of $\mathbf{y}$ that satisfy (2.225).

On the other hand, if in addition to knowing $\mathbf{z}$, $\mathbf{x}$ is also known, then the sample space of $\mathbf{y}$ is a singleton. That is, $\underline{\mathbf{y}}$ can be determined uniquely.

The condition (2.225) occurs only if the function $\mathbf{f}$ is non-invertible.

□

Chapter 3

Distributions

3.1 Bernoulli distribution

From [1].

Definition 3.1.1 (*Bernoulli drandom variable*)**.** A random variable $\underline{x}$ is called Bernoulli with parameter p if there exists an event A with probability $p = P(A)$ such that

$$x(s) = \begin{cases} 1, & s \in A \\ 0, & s \notin A, \end{cases} \tag{3.1}$$

where $s \in S$.

Remark 3.1.1. Some remarks on Bernoulli random variables.

- $\underline{x}$ is a *discrete* random variable with range $\mathcal{R}_x = \{0, 1\}$.
- PMF of $\underline{x}$

$$p(0) = P(\underline{x} = 0) = P\left(A^{\mathrm{C}}\right) = 1 - p = q \tag{3.2}$$
$$p(1) = P(\underline{x} = 1) = P(A) = p \tag{3.3}$$
$$p(x) = 0 \qquad \text{if } x \notin \{0, 1\}. \tag{3.4}$$

3.2 Geometric random variables

From [1].

Definition 3.2.1 (*Geometric random variable*)**.** A geometric random variable is a discrete random variable with a countable infinite set of possible values. The set of possible values for the geometric random

variable $\underline{x}$ is

$$\mathcal{R}_x = \{1, 2, \ldots\} = \mathbb{N}. \tag{3.5}$$

Theorem 3.2.1 (PMF of geometric random variables)**.** Let $\underline{x}$ be a geometric random variable with parameter p. The PMF of $\underline{x}$ is then

$$p(x) = \begin{cases} (1-p)^{x-1} p, & x = 1, 2, \ldots, \\ 0, & \text{otherwise.} \end{cases} \tag{3.6}$$

3.3 Binomial random variable

Definition 3.3.1 (*Binomial random variable*)**.** Consider a sequence of n identical and independent Bernoulli trials with probability of success p. The random variable $\underline{x}$ defined by

$$\underline{x} = \text{number of successes in the } n \text{ trials} \tag{3.7}$$

is called binomial with parameters n and p, or simply $B(n, p)$.

Remark 3.3.1. Some remarks on Binomial random variables.

- Set of possible values of $\underline{x}$:

$$\mathcal{R}_x = \{0, 1, \ldots, n\}. \tag{3.8}$$

- Notation:

$$\underline{x} \sim B(n, p). \tag{3.9}$$

- Basic examples
  - Number of heads in a sequence of 10 independent flips of a coin.
  - Number of defective IC chips in a production sample of size n.

Theorem 3.3.1 (PMF of a Binomial random variable)**.** Let $\underline{x} \sim B(n, p)$. The PMF of $\underline{x}$ is

$$p(x) = \begin{cases} \binom{n}{k} p^x (1-p)^{n-x}, & x = 0, 1, \ldots, n \\ 0, & \text{otherwise.} \end{cases} \tag{3.10}$$

3.4 Poisson distribution

A discrete random variable $\underline{x}$ is called Poisson with parameter $\lambda > 0$ if its PMF has the form

$$p(x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda}, & x \in \mathbb{N}_0 \\ 0, & \text{otherwise,} \end{cases} \tag{3.11}$$

where $\mathbb{N}_0 = \{0, 1, \ldots\}$.

3.5 Uniform random variables

Definition 3.5.1 (*Uniform random variable*)**.** A continuous random variable $\underline{x}$ is called uniform over the interval (a, b), written as $\underline{x} \sim U(a, b)$, where $a < b$ are real numbers, if its PDF is

$$f(x) = \begin{cases} \frac{1}{b-a}, & a < x < b, \\ 0, & \text{otherwise.} \end{cases} \tag{3.12}$$

Theorem 3.5.1 (Unoform random variable CDF)**.** The CDF is given by

$$F(x) = \begin{cases} 0, & x \leq a, \\ \frac{x-a}{b-a}, & a \leq x \leq b, \\ 1 & x \geq b. \end{cases} \tag{3.13}$$

3.6 Exponential random variable

Definition 3.6.1 (*Exponential random variable*)**.** A continuous random variable $\underline{x}$ is called exponential with parameter $\lambda > 0$ if its PDF is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{3.14}$$

Theorem 3.6.1 (Exponential random variable CDF)**.** The CDF is given by

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & x > 0, \\ 0, x \leq 0. \end{cases} \tag{3.15}$$

3.7 Gaussian random variables

Definition 3.7.1 (*Gaussian (normal) distribution*)**.** A random vector $\underline{\mathbf{x}}$ is called a *Gaussian (or normal) random vector* with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$ if it follows a *Gaussian (or normal) distribution*

given by

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det \mathbf{C}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{3.16}$$

$$= \mathcal{N}(\boldsymbol{\mu}, \mathbf{C}). \tag{3.17}$$

The n elements of $\mathbf{x}$, $\underline{x}_1, \ldots, \underline{x}_n$ are called *jointly normal random variables*.

Definition 3.7.2 (*Standard Normal Distribution*)**.** A *standard normal distribution* is a normal distribution with a special case of $\boldsymbol{\mu} = \mathbf{0}$ and $\mathbf{C} = \mathbf{1}$. That is,

$$f(\mathbf{x}) = \mathcal{N}(\mathbf{0}, \mathbf{1}). \tag{3.18}$$

3.7.1 Properties of Gaussian random variables

Theorem 3.7.1 (Linear combination of normal RVs)**.** A random vector $\underline{\mathbf{x}}$ is called a Gaussian (or normal) random vector if and only if the random variable $\mathbf{a}^\mathsf{T}\underline{\mathbf{x}}$ is a Gaussian random variable for any vector $\mathbf{a}$.

Remark 3.7.1 (Properties of Gaussian random variables)**.** Here are some properties of Gaussian random variables.

1. A Gaussian random vector consists of marginally Gaussian random variables.
2. However, marginally Gaussian random variables are *not* necessarily jointly Gaussian.[a]

[a] That is, if you take 3 Gaussian random variables, then they are not necessarily jointly Gaussian. Unless they are independent.

Definition 3.7.3 (*Complex Gaussian radnom vector*)**.** A complex random vector

$$\underline{\mathbf{z}} = \begin{bmatrix} \underline{x}_1 + \jmath \underline{y}_1 \\ \vdots \\ \underline{x}_n + \jmath \underline{y}_n \end{bmatrix} \tag{3.19}$$

is a *complex Gaussian random vector* if and only if the real random variables $\underline{x}_1, \underline{y}_1, \ldots, \underline{x}_n, \underline{y}_n$ are jointly normal.

Theorem 3.7.2 (Characteristic function of normal random variables)**.** If the random variables $\underline{x}_1, \ldots, \underline{x}_n$ are jointly Normal with mean 0 and covariance matrix $\mathbf{C}$. Then, their characteristic function is

$$\Phi_{\mathbf{x}}(\Omega) = e^{-\frac{1}{2}\Omega^\mathsf{T}\mathbf{C}\Omega}. \tag{3.20}$$

3.8 Chi-squared distribution

Definition 3.8.1 (*Chi-squared distribution*)**.** If $\underline{x}_1, \ldots, \underline{x}_n$ are ***independent*** *standard normal variables* (check Definition 3.7.2), then the sum of their squares

$$\underline{y} = \sum_{k=1}^{n} \underline{x}_k^2 \in \mathbb{R} \tag{3.21}$$

is distributed according to the Chi-square distribution with n degrees of freedom, denoted by χ_n^2. That is,

$$\underline{y} \sim \chi_n^2. \tag{3.22}$$

Remark 3.8.1. If we have a random vector $\underline{\mathbf{x}} \in \mathbb{R}^n$ with mean $\boldsymbol{\mu}_x = \mathbf{0}$ and a unit covariance $\mathbf{C} = \mathbf{1}$, then the sum of squares

$$\underline{y} = \underline{\mathbf{x}}^{\mathsf{T}} \underline{\mathbf{x}} \tag{3.23}$$

$$\tag{3.24}$$

follows a Chi-squared distribution with n degrees of freedom. That is

$$\underline{y} \sim \chi_n^2. \tag{3.25}$$

Remark 3.8.2. MATLAB has a built-in inverse CDF function called `chi2inv(p,nu)` that evaluates the inverse CDF of a Chi-squared distribution at the probability value $p \in [0, 1]$ and the degrees of freedom ν (which we denote as n in Definition 3.8.1).

3.8.1 Relation to the Mahalanobis distance

In estimation problems, the Mahalanobis distance is often encountered. Especially when testing for consistency of the estimates.

Definition 3.8.2 (*Mahalanobis distance*)**.** The *Mahalanobis distance* $D_M \in \mathbb{R}$ is defined as

$$D_M \triangleq \sqrt{\mathbf{r}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \mathbf{r}}, \tag{3.26}$$

where $\mathbf{r} \in \mathbb{R}^n$ is a sample of

$$\underline{\mathbf{r}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{3.27}$$

It is a *Euclidean distance* normalized by *uncertainty*.

Example 3.8.1. Consider a random variable $\underline{\mathbf{x}} \in \mathbb{R}^n$ distributed according to $\underline{\mathbf{x}} \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$. Say a

realization $\mathbf{x} \in \mathbb{R}^n$ is sampled from the distribution. Then the Mahalanobis distance

$$D_M = \sqrt{(\mathbf{x} - \boldsymbol{\mu}_x)^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})} \tag{3.28}$$

is a measure of how far off the realization $\mathbf{x}$ is from the mean $\boldsymbol{\mu}_x$. △

Theorem 3.8.1. The squared Mahalanobis distance follows a Chi-squared distribution. That is,

$$\underline{D_M}^2 \sim \chi^2_n. \tag{3.29}$$

Proof. Given a random variable $\underline{\mathbf{r}} \in \mathbb{R}^n$ that follows the distribution $\underline{\mathbf{r}} \sim (\mathbf{0}, \boldsymbol{\Sigma})$, then the Mahalanobis distance squared is a random variable. Specifically,

$$\underline{D}_M^2 = \underline{\mathbf{r}}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \underline{\mathbf{r}}. \tag{3.30}$$

Since the covariance matrix $\boldsymbol{\Sigma}$ is symmetric positive definite, then its eigenvectors are orthonormal and the eigenvalues are positive. Therefore, (3.30) can be written as

$$\begin{aligned}
\underline{D}_M^2 &= \underline{\mathbf{r}}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \underline{\mathbf{r}} && (3.31)\\
&= \underline{\mathbf{r}}^\mathsf{T} \left(\mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^\mathsf{T} \right)^{-1} \underline{\mathbf{r}} && (3.32)\\
&= \underline{\mathbf{r}}^\mathsf{T} \mathbf{U} \boldsymbol{\Lambda}^{-1} \mathbf{U}^\mathsf{T} \underline{\mathbf{r}} && (3.33)\\
&= \underline{\mathbf{r}}^\mathsf{T} \mathbf{U} \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{\Lambda}^{-\frac{1}{2}} \mathbf{U}^\mathsf{T} \underline{\mathbf{r}} && (3.34)\\
&= \left(\boldsymbol{\Lambda}^{-\frac{1}{2}} \mathbf{U}^\mathsf{T} \underline{\mathbf{r}} \right)^\mathsf{T} \left(\boldsymbol{\Lambda}^{-\frac{1}{2}} \mathbf{U}^\mathsf{T} \underline{\mathbf{r}} \right) && (3.35)\\
&= \underline{\mathbf{y}}^\mathsf{T} \underline{\mathbf{y}}, && (3.36)
\end{aligned}$$

where

$$\underline{\mathbf{y}} = \boldsymbol{\Lambda}^{-\frac{1}{2}} \mathbf{U}^\mathsf{T} \underline{\mathbf{r}} \in \mathbb{R}^n. \tag{3.37}$$

The next step is to check whether $\underline{\mathbf{y}}$ is a standard normal variable. First, let's check the mean.

$$\begin{aligned}
\boldsymbol{\mu}_y &= \mathbb{E}\left[\underline{\mathbf{y}}\right] && (3.38)\\
&= \boldsymbol{\Lambda}^{-\frac{1}{2}} \mathbf{U}^\mathsf{T} \underbrace{\mathbb{E}\left[\underline{\mathbf{r}}\right]}_{\mathbf{0}} && (3.39)\\
&= \mathbf{0}. && (3.40)
\end{aligned}$$

The second condition to check is whether the covariance of $\underline{\mathbf{y}}$ is the identity matrix or not.

$$\begin{aligned}
\mathrm{Cov}\left[\underline{\mathbf{y}}\right] &= \mathbb{E}\left[\underline{\mathbf{y}}\underline{\mathbf{y}}^{\mathsf{T}}\right] && (3.41)\\
&= \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}^{\mathsf{T}}\underbrace{\mathbb{E}\left[\underline{\mathbf{r}}\underline{\mathbf{r}}^{\mathsf{T}}\right]}_{\mathbf{\Sigma}}\mathbf{U}\mathbf{\Lambda}^{-\frac{1}{2}} && (3.42)\\
&= \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}^{\mathsf{T}}\underbrace{\mathbf{\Sigma}}_{\mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathsf{T}}}\mathbf{U}\mathbf{\Lambda}^{-\frac{1}{2}} && (3.43)\\
&= \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{U}^{\mathsf{T}}\cancelto{\mathbf{1}}{\mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathsf{T}}}\cancelto{\mathbf{1}}{\mathbf{U}}\mathbf{\Lambda}^{-\frac{1}{2}} && (3.44)\\
&= \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{\Lambda}\mathbf{\Lambda}^{-\frac{1}{2}} && (3.45)\\
&= \mathbf{1}. && (3.46)
\end{aligned}$$

Thus, $\underline{\mathbf{y}}$ is a standard normal random variable. Therefore, according to Definition 3.8.1, $\underline{D}_M^2$ follows a Chi-squared distribution with n degrees of freedom. □

Chapter 4

Graphs in probability

4.1 Graph definitions and terminology

Definition 4.1.1 (*Directed graphs*)**.** A directed graph consists of a set of vertices and a set of directed edges that connect pairs of vertices.

In probability, the vertices are random variables. Figure 4.1 shows an example of a directed graph.

Definition 4.1.2 (*Directed acyclic graph*)**.** A directed acyclic graph (DAG) is a directed graph that contains no cycles. That is, no directed paths from a vertex back to itself.

In case of probability, a DAG is called a *Bayesian network* (BN).

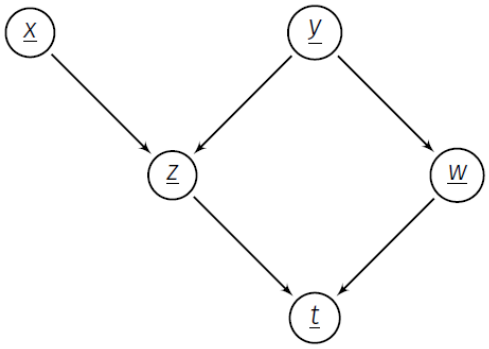

Figure 4.1: Illustration of a simple directed graph in probability.

The terms are illustrated in the figures below.

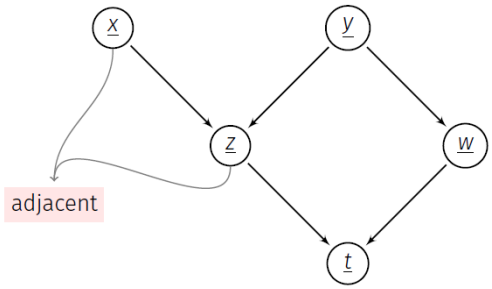

Figure 4.2: Directed graph terminology: adjacent vertices.

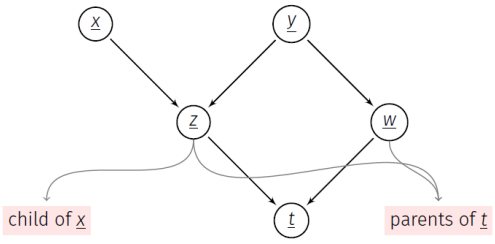

Figure 4.3: Directed graph terminology: child and parent vertices.

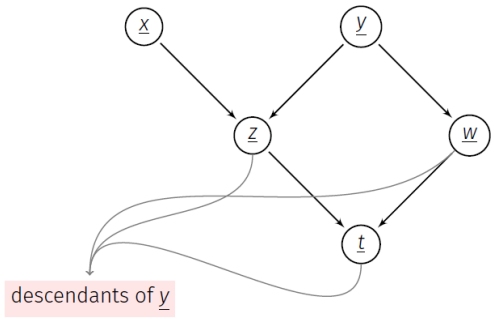

Figure 4.4: Directed graph terminology: descendants vertices.

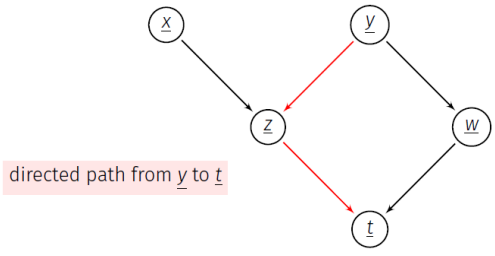

Figure 4.5: Directed graph terminology: directed path.

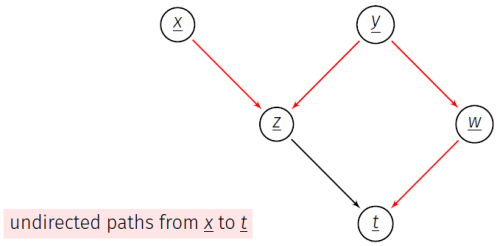

Figure 4.6: Directed graph terminology: undirected path.

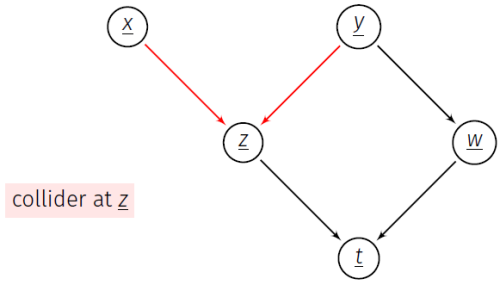

Figure 4.7: Directed graph terminology: collider.

4.2 Graphs in Probability

Definition 4.2.1 (*Bayesian networks*)**.** A Bayesian network is a directed acyclic graph (DAG) (check Def. 4.1.2) that has random variables as vertices.

Theorem 4.2.1 (Representing JPDFs using DAGs)**.** Representing JPDFs using DAGs Consider n random variables $\underline{x}_1, \ldots, \underline{x}_n$ with joint pdf (JPDF) $f(x_1, \ldots, x_n)$. Let G be a DAG with n vertices representing the random variables. Then, G represents the JPDF f if

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i | \text{ parents of } x_i). \tag{4.1}$$

Theorem 4.2.2 (Markov condition in DAG)**.** The DAG G represents the JPDF f if and only if the Markov condition holds, which states the following:

$$\forall \underline{x}_i : \quad \underline{x}_i \perp \bar{\mathcal{X}}_i | \text{parents of } \underline{x}_i, \tag{4.2}$$

where $\bar{\mathcal{X}}_i$ are all other random variables but the parents and descendants of $\underline{x}_i$.

Definition 4.2.2 (*D-donnection and d-separation*)**.** Consider a set of n random variables $\underline{x}_1, \ldots, \underline{x}_n$ whose JPDF is represented by G. Furthermore, let $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ be mutually exclusive sets with $\mathcal{A}, \mathcal{C} \neq \emptyset$.

The sets $\mathcal{A}$ and $\mathcal{C}$ are called *d-connected* given $\mathcal{B}$ if there exists an *undirected* path between some random

variable in $\mathcal{A}$ and some random variable in $\mathcal{C}$ such that

1. every collider along the path is in $\mathcal{B}$, or has a descendent in $\mathcal{B}$;
2. no non-collider along the path belongs in $\mathcal{B}$.

If $\mathcal{A}$ and $\mathcal{C}$ are not d-connected given $\mathcal{B}$, then they are called *d-separated given* $\mathcal{B}$.

Theorem 4.2.3 (D-connection and d-separation)**.** Consider a set of n random variables $\underline{x}_1, \ldots, \underline{x}_n$ whose JPDF is represented by G. Furthermore, let $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ be mutually exclusive sets with $\mathcal{A}, \mathcal{C} \neq \emptyset$. Then, the sets $\mathcal{A}$ and $\mathcal{C}$ are *d-separated given* $\mathcal{B}$ if and only if

$$\mathcal{A} \perp \mathcal{C} | \mathcal{B}. \tag{4.3}$$

Part II

Statistical Inference

Chapter 5

Parameter Estimation

5.1 Motivation

Let $\underline{x}$ be a random variable whose CDF, $F(x)$, *is unknown*. The goal is estimate $F(x)$ from n observations of $\underline{x}$, $x_1, \ldots, x_n$.

Definition 5.1.1 (*Empirical CDF*)**.** The empirical CDF of $\underline{x}$ from the observations $x_1, \ldots, x_n$ is

$$\hat{F}(x) = \frac{1}{n} \times (\#\text{ of observations } x_i \leq x)\,. \tag{5.1}$$

The empirical CDF $\hat{F}(x)$ is the CDF of a discrete random variable with PMF

$$\hat{p}_n(x) = \begin{cases} \frac{1}{n}, & x = x_i, \quad i = 1, \ldots, n, \\ 0 & \text{Otherwise.} \end{cases} \tag{5.2}$$

The PDF is then given by

$$\hat{f}_n(x) = \sum_{i=1}^{n} \frac{1}{n} \delta(x - x_i). \tag{5.3}$$

The empirical PDF $\hat{f}(x)$ in (5.3) assigns constant weights to every observation. The empirical PDF can be generalized by giving different weights to different observations. Specifically,

$$\hat{f}_n(x) = \sum_{i=1}^{n} \frac{1}{n} k(x - x_i), \tag{5.4}$$

where $k(\cdot)$ is the kernel function such that it is

1. symmetric around 0, and
2. $\int_{-\infty}^{\infty} k(x)\mathrm{d}x = 1$.

A common choice is the Gaussian kernel. That is, a PDF of $\mathcal{N}(0, 1)$.

5.2 Problem statement

Let $\underline{\mathbf{x}}$ be a random variable. Furthermore, let $f(\mathbf{x}; \boldsymbol{\theta})$ be the PDF of $\underline{\mathbf{x}}$ of *known form* (e.g., Gaussian distribution) but depending on *unknown* parameters $\boldsymbol{\theta}$ (e.g., mean and covariance of a Gaussian distribution).

Example 5.2.1. A Gaussian random variable has a PDF parameterized w.r.t. to the mean μ and the variance σ^2. Thus, the parameters could be $\boldsymbol{\theta} = \left[\mu, \sigma^2\right]$.

The goal is to estimate $\boldsymbol{\theta}$ from n observations of $\underline{\mathbf{x}}$: $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Define the data vector

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{bmatrix}. \tag{5.5}$$

The data vector $\mathbf{X}$ is a realization of the random vector

$$\underline{\mathbf{X}} = \begin{bmatrix} \underline{\mathbf{x}}_1 \\ \vdots \\ \underline{\mathbf{x}}_n \end{bmatrix}, \tag{5.6}$$

where $\underline{\mathbf{x}}_i$, $i = 1, \ldots, n$ are i.i.d.

Approaches

There are two approaches to parameter estimation.

1. *Point estimation*: identify a function $g(\cdot)$ of the sample such that $g(\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_n)$ is a good (in some sense) approximation of $\boldsymbol{\theta}$.
2. *Interval estimation*: identify two functions $g_1(\cdot)$ and $g_2(\cdot)$ of the sample, such that

$$P\left(g_1(\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_n) < \boldsymbol{\theta} < g_2(\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_n)\right) = \gamma, \tag{5.7}$$

   where γ is the confidence interval.

5.3 Definitions and terminology

Definition 5.3.1 (*Estimator*)**.** The estimator of $\boldsymbol{\theta}$ is given in the form

$$\underline{\hat{\boldsymbol{\theta}}} = g(\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_n). \tag{5.8}$$

The estimator models the behaviour of the estimation method over all possible samples.

Definition 5.3.2 (*Estimate of $\boldsymbol{\theta}$*)**.** The *deterministic* quantity $g(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. That is, the the output of $\underline{\hat{\boldsymbol{\theta}}}$ for the realizations $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Definition 5.3.3 (*Statistic*)**.** Any function of $\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_n$.

Definition 5.3.4 (*Bias*)**.** The estimator $g(\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_n)$ is called

- *unbiased* if

$$\mathbb{E}\left[g(\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_n)\right] = \boldsymbol{\theta}. \tag{5.9}$$

- Otherwise, it's called *biased* with bias

$$\mathbb{E}\left[g(\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_n)\right] - \boldsymbol{\theta}. \tag{5.10}$$

Definition 5.3.5 (*Asymptic unbias*)**.** A biased estimator for which

$$\mathbb{E}\left[g(\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_n)\right] \overset{N\to\infty}{\to} \boldsymbol{\theta} \tag{5.11}$$

is called *asymptotically unbiased*.

Definition 5.3.6 (*Mean-squared best estimator*)**.** For a fixed n, the estimator that minimizes the mean-squared (MS) error

$$\mathbf{e}_n = \mathbb{E}\left[\left|g(\underline{\mathbf{x}}_1, \ldots \underline{\mathbf{x}}_n)^{\mathsf{T}} g(\underline{\mathbf{x}}_1, \ldots \underline{\mathbf{x}}_n)\right|\right] \tag{5.12}$$

is called the *best* estimator (in the mean-squared sense).

Definition 5.3.7 (*Consistency*)**.** The estimator $g(\underline{\mathbf{x}}_1, \ldots \underline{\mathbf{x}}_n)$ is called *consistent* if

$$\lim_{n\to\infty} P\left(|g(\underline{\mathbf{x}}_1, \ldots \underline{\mathbf{x}}_n)| > \boldsymbol{\epsilon}\right) = 0, \forall \boldsymbol{\epsilon} > \mathbf{0}, \tag{5.13}$$

or

$$\lim_{n\to\infty} P\left(|g(\underline{\mathbf{x}}_1, \ldots \underline{\mathbf{x}}_n)| < \boldsymbol{\epsilon}\right) = 1, \forall \boldsymbol{\epsilon} > \mathbf{0}. \tag{5.14}$$

Theorem 5.3.1 (Sufficient condition for consistency)**.** Let $g(\underline{\mathbf{x}}_1, \ldots \underline{\mathbf{x}}_n)$ be an estimator of $\boldsymbol{\theta}$. Then, if

$$\lim_{n\to\infty} \mathbb{E}\left[\left|g(\underline{\mathbf{x}}_1, \ldots \underline{\mathbf{x}}_n)^{\mathsf{T}} g(\underline{\mathbf{x}}_1, \ldots \underline{\mathbf{x}}_n)\right|\right] = \mathbf{0}, \tag{5.15}$$

then $g(\underline{\mathbf{x}}_1, \ldots \underline{\mathbf{x}}_n)$ is a *consistent estimator* of $\boldsymbol{\theta}$.

The converse is not true in general. That is, this is a sufficient but not necessary condition.

Theorem 5.3.2 (Sample mean estimator)**.** Let $\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_n$ be a random sample of size n and let $\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_n$ be *uncorrelated*. The *sample mean estimator*

$$\underline{\hat{\boldsymbol{\mu}}}_n = \frac{1}{n} \sum_{i=1}^{n} \underline{\mathbf{x}}_i \tag{5.16}$$

is

- an *unbiased estimator* of $\boldsymbol{\mu}$;
- a *consistent estimator* of $\boldsymbol{\mu}$.

Theorem 5.3.3 (Variance estimaton)**.** Consider the sample average estimators of the mean and variance of Gaussian random variable $\underline{x}$ from a random sample $\underline{x}_1, \ldots, \underline{x}_n$ of size n:

$$\underline{\hat{\mu}}_n = \frac{1}{n} \sum_{i=1}^{n} \underline{x}_i, \tag{5.17}$$

$$\underline{\hat{\sigma}}_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\underline{x}_i - \underline{\hat{\mu}}_n \right)^2. \tag{5.18}$$

The random variables $\underline{\hat{\mu}}_n$ and $\underline{\hat{sigma}}_n^2$ are independent.

5.4 Method of moments

Recall that moments are defined in Def. 1.6.4 and Def. 2.4.1. The m-th moment estimator is given by

$$\hat{m}_m(\mathbf{X}) = \frac{1}{n} \sum_{i}^{n} x_i^m. \tag{5.19}$$

Let $\boldsymbol{\theta} \in \mathbb{R}^k$ be of size k and let

$$\mathbf{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \tag{5.20}$$

be the samples of the random variable $\underline{x}$. Then, the idea of method of moments is to estimate the first k moments, thus getting k equations, and then solving the k equations for $\boldsymbol{\theta}$.

Example 5.4.1. Estimate the mean μ and variance σ^2 of a random variable $\underline{x}$ from a random sample $\underline{\mathbf{X}} = \begin{bmatrix} \underline{x}_1 & \cdots & \underline{x}_n \end{bmatrix}^\mathsf{T}$.

The moments are given by

$$m_1(\mu, \sigma^2) = \mathbb{E}[\underline{x}] \quad = \mu, \tag{5.21}$$
$$m_2(\mu, \sigma^2) = \mathbb{E}\left[\underline{x}^2\right] = \mu^2 + \sigma^2. \tag{5.22}$$

The sample average estimates of the moments are given by

$$\hat{m}_1(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{5.23}$$
$$\hat{x}_2(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} x_i^2. \tag{5.24}$$

Thus, the parameter estimates can be computed by solving the system of equations

$$\mu = \hat{m}_1(\mathbf{X}) \tag{5.25}$$
$$\mu^2 + \sigma^2 = \hat{m}_2(\mathbf{X}). \tag{5.26}$$

The solution is given by

$$\hat{\mu} = \hat{m}_1(\mathbf{X}) \tag{5.27}$$
$$= \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{5.28}$$
$$\hat{\sigma}^2 = \hat{m}_2(\mathbf{X}) - \hat{m}_1^2(\mathbf{X}) \tag{5.29}$$
$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2. \tag{5.30}$$

5.5 Maximum likelihood (ML) estimator

Definition 5.5.1 (*Likelihood*)**.** Given the observations $\underline{\mathbf{x}}_1 = \mathbf{x}_1', \ldots, \underline{\mathbf{x}}_n = \mathbf{x}_n'$, the *likelihood* of $\boldsymbol{\theta}$ is the joint PDF of $\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_n$ evaluated at the observed points

$$L(\boldsymbol{\theta}; \mathbf{x}_1', \ldots, \mathbf{x}_N') = f\left(\mathbf{x}_1', \ldots, \mathbf{x}_n'; \boldsymbol{\theta}\right) \tag{5.31}$$
$$= f_{\mathbf{x}}(\mathbf{x}_1'; \boldsymbol{\theta}) \cdots f_{\mathbf{x}}(\mathbf{x}_n'; \boldsymbol{\theta}). \tag{5.32}$$

Remark 5.5.1.

- The likelihood $f\left(\mathbf{x}_1', \ldots, \mathbf{x}_n'; \boldsymbol{\theta}\right)$ is treated as a function of $\boldsymbol{\theta}$.
- It is not a PDF. That is,

$$\int_{-\infty}^{\infty} f\left(\mathbf{x}_1', \ldots, \mathbf{x}_n'; \boldsymbol{\theta}\right) \mathrm{d}\boldsymbol{\theta} \neq 1. \tag{5.33}$$

Definition 5.5.2 (*Maximum-likelihood estimator*)**.** The maximum-likelihood (ML) estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$, is the value of $\boldsymbol{\theta}$ that maximizes $f\left(\mathbf{x}'_1, \ldots, \mathbf{x}'_n; \boldsymbol{\theta}\right)$. That is,

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^k} L(\boldsymbol{\theta}; \mathbf{x}'_1, \ldots, \mathbf{x}'_N) \tag{5.34}$$

$$= \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^k} f\left(\mathbf{x}'_1, \ldots, \mathbf{x}'_n; \boldsymbol{\theta}\right). \tag{5.35}$$

Remark 5.5.2. Some remarks regarding the ML estimator.

- Log-likelihood function

$$\ell(\mathbf{x}'_1, \ldots, \mathbf{x}'_n; \boldsymbol{\theta}) = \log L(\boldsymbol{\theta}; \mathbf{x}'_1, \ldots, \mathbf{x}'_N) \tag{5.36}$$

$$= \log f\left(\mathbf{x}'_1, \ldots, \mathbf{x}'_n; \boldsymbol{\theta}\right). \tag{5.37}$$

- The maximizer of the log-likelihood function is the same maximizer of the likelihood function. That is,

$$\arg\max_{\boldsymbol{\theta} \in \mathbb{R}^k} L(\boldsymbol{\theta}; \mathbf{x}'_1, \ldots, \mathbf{x}'_N) = \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^k} \ell(\boldsymbol{\theta}; \mathbf{x}'_1, \ldots, \mathbf{x}'_N). \tag{5.38}$$

- The ML estimate is

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^k} f\left(\mathbf{x}'_1, \ldots, \mathbf{x}'_n; \boldsymbol{\theta}\right) \tag{5.39}$$

$$= \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^k} \log f\left(\mathbf{x}'_1, \ldots, \mathbf{x}'_n; \boldsymbol{\theta}\right) \tag{5.40}$$

$$= \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^k} L(\mathbf{x}'_1, \ldots, \mathbf{x}'_n; \boldsymbol{\theta}). \tag{5.41}$$

Remark 5.5.3. Some remarks on ML estimation.

1. Let $\underline{\hat{\boldsymbol{\theta}}}_{\mathrm{ML}}$ be the ML estimator of $\boldsymbol{\theta}$ and let $g(\boldsymbol{\theta})$ be a function of $\boldsymbol{\theta}$. Then, the ML estimator of $g(\boldsymbol{\theta})$ is $g(\underline{\hat{\boldsymbol{\theta}}}_{\mathrm{ML}})$.
2. Limitations of the ML estimation:
   2.1. The ML estimate may not exist.
   2.2. The ML estimate may not be unique.
3. Maximum likelihood estimate can be finding by minimizing the negative likelihood or negative log-likelihood.

Example 5.5.1. Let $\underline{x} \sim U[0, \theta]$ be a uniformly distributed random variable in the form

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}, & x \in [0,\theta], \\ 0, & \text{otherwise}, \end{cases} \tag{5.42}$$

where θ is some unknown parameter. Find the ML estimator of θ from a random sample of size N.

1. First, get the joint PDF of all the random samples

$$f(x_1,\ldots,x_N;\theta) = f(x_1;\theta)\cdots f(x_N;\theta) \tag{5.43}$$

$$= \begin{cases} \frac{1}{\theta^N}, & x_i \in [0,\theta], i = 1,\ldots,N, \\ 0, & \text{otherwise}, \end{cases} \tag{5.44}$$

   where the fact that the random samples $\underline{x}_i$ for $i = 1,\ldots,N$ are i.i.d.

2. The likelihood of θ is then given as a function of the samples of $\underline{x}_i = x_i'$. Specifically, the likelihood is

$$L(\theta; x_1',\ldots,x_N') = f\left(\underline{x}_1 = x_1',\ldots,\underline{x}_N = x_N';\theta\right) \tag{5.45}$$

$$= \begin{cases} \frac{1}{\theta^N}, & x_i' \in [0,\theta], i = 1,\ldots,N, \\ 0, & \text{otherwise} \end{cases} \tag{5.46}$$

$$= \begin{cases} \frac{1}{\theta^N}, & \max_i(x_i) \le \theta, \\ 0, & \text{otherwise}. \end{cases} \tag{5.47}$$

   Note that it was not necessary to specify the lower bound $\min_i(x_i) \ge 0$ since the samples are drawn from a uniform distribution $U[0,\theta]$ which is bounded from below by 0. Thus, $P(\underline{x} < 0) = 0$. Therefore, it is impossible to have a sample $x_i' < 0$.

3. Maximize the likelihood function. Maximizing (5.47) is a *constrained* optimization problem that entails maximizing $\frac{1}{\theta^N}$.

   Maximizing $\frac{1}{\theta^N}$ is the same as minimizing θ. However, since θ is bounded from below by the constraint in (5.47), then the minimum value of θ is $\max_i(x_i)$. Thus,

$$\hat{\theta}_{\text{ML}} = \max_i(x_i). \tag{5.48}$$

△

5.6 Bayesian parameter estimation

In the previous sections, the approaches used were the *classical* or *frequentist* approaches. The classical approach

1. treats the unknown parameter as deterministic parameter, and

2. assumes no prior information on the unknown parameter was available.

For Bayesian approach on the other hand

1. assumes that unknown parameter is random with known pdf $f(\boldsymbol{\theta})$, called the *prior*;
2. the JPDF of random sample given $\underline{\boldsymbol{\theta}} = \theta$ (likelihood) is given by

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_n | \boldsymbol{\theta}); \tag{5.49}$$

3. The estimation problem becomes a prediction problem; predict $\underline{\boldsymbol{\theta}}$ from a correlated (with $\underline{\boldsymbol{\theta}}$) sample $\underline{\mathbf{X}}$.

The mean-squared (or Bayes) estimate $\hat{\boldsymbol{\theta}}$ of $\underline{\boldsymbol{\theta}}$ given $\underline{\mathbf{X}} = \mathbf{X}$ is

$$\hat{\boldsymbol{\theta}} = \mathbb{E}[\underline{\boldsymbol{\theta}} | \mathbf{X}] \tag{5.50}$$

$$= \int_{-\infty}^{\infty} \boldsymbol{\theta} f(\boldsymbol{\theta} | \mathbf{X}) \, \mathrm{d}\boldsymbol{\theta}, \tag{5.51}$$

where

$$f(\boldsymbol{\theta} | \mathbf{X}) = \frac{f(\mathbf{X} | \boldsymbol{\theta}) f(\boldsymbol{\theta})}{f(\mathbf{X})} \tag{5.52}$$

$$= \frac{f(\mathbf{X} | \boldsymbol{\theta}) f(\boldsymbol{\theta})}{\int_{-\infty}^{\infty} f(\mathbf{X} | \boldsymbol{\theta}) f(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}}. \tag{5.53}$$

Terminology

- The mean-squared estimator is also referred to as the Bayes estimator.
- $f(\boldsymbol{\theta})$: *prior* pdf of $\underline{\boldsymbol{\theta}}$ (before observing the sample).
- $f(\boldsymbol{\theta} | \mathbf{X})$: *posterior* pdf of $\boldsymbol{\theta}$ (after observing the sample).
- In [5], the term $f(\mathbf{X} | \boldsymbol{\theta})$ is sometimes referred to as the likelihood of $\mathbf{X}$ given $\boldsymbol{\theta}$. This is not to be confused with the likelihood from the ML estimator. In estimation, the term $f(\mathbf{X} | \boldsymbol{\theta})$ can also be referred to as an *observation model*.

Remark 5.6.1. Some remarks about Bayesian estimation.

- $\mathbb{E}[\underline{\boldsymbol{\theta}} | \mathbf{X}]$ from (5.50) is difficult to solve for two reasons. First, the denominator of the posterior pdf involves evaluating an integral over the entire sample space of $\boldsymbol{\theta}$. Second, evaluating the expectation $\mathbb{E}[\cdot]$ requires *another* integral to be evaluated over the entire sample space.
- Maximum a-posteriori (MAP) estimate maximizes the *posterior* pdf $f(\boldsymbol{\theta} | \mathbf{X})$ with respect to $\boldsymbol{\theta}$.

Since the denominator of the prior is constant, then it can be omitted. That is,

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} f\left(\boldsymbol{\theta}|\, \mathbf{X}\right) \tag{5.54}$$

$$= \arg\max_{\boldsymbol{\theta}} \frac{f\left(\mathbf{X}|\, \boldsymbol{\theta}\right) f\left(\boldsymbol{\theta}\right)}{f\left(\mathbf{X}\right)} \tag{5.55}$$

$$= \arg\max_{\boldsymbol{\theta}} f\left(\mathbf{X}|\, \boldsymbol{\theta}\right) f\left(\boldsymbol{\theta}\right). \tag{5.56}$$

5.7 Maximum a-posteriori (MAP) estimator

This estimator is an *approximation* to the Bayes estimator discussed in the previous section. It is easier (less computationally intensive) to compute.

Definition 5.7.1 (*Maximum a-posteriori estimator*)**.** Maximum a-posteriori (MAP) estimate maximizes the *posterior* pdf $f\left(\boldsymbol{\theta}|\, \mathbf{X}\right)$ with respect to $\boldsymbol{\theta}$. Since the denominator of the prior is constant, then it can be omitted. That is,

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} f\left(\boldsymbol{\theta}|\, \mathbf{X}\right) \tag{5.57}$$

$$= \arg\max_{\boldsymbol{\theta}} \frac{f\left(\mathbf{X}|\, \boldsymbol{\theta}\right) f\left(\boldsymbol{\theta}\right)}{f\left(\mathbf{X}\right)} \tag{5.58}$$

$$= \arg\max_{\boldsymbol{\theta}} f\left(\mathbf{X}|\, \boldsymbol{\theta}\right) f\left(\boldsymbol{\theta}\right). \tag{5.59}$$

5.7.1 MAP estimator of a Markov normally distributed random variable

In this section, the MAP estimator for normally distributed variables that follow a Markov chain will be derived. For each problem, a process model is provided in the form

$$\underline{\mathbf{x}}_{k+1} = \underline{\mathbf{f}}(\mathbf{x}_k, \underline{\mathbf{u}}_k, \underline{\mathbf{w}}_k), \tag{5.60}$$

where $\underline{\mathbf{w}}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_k)$ is the process noise and the covariance matrix $\mathbf{Q}_K$ is known. A prior is distributed according to $\hat{\mathbf{x}}_0 \sim \mathcal{N}\left(\hat{\mathbf{x}}_1, \check{\mathbf{P}}_0\right)$.

Furthermore, a measurement model is given in the form

$$\underline{\mathbf{y}}_k = \mathbf{g}(\underline{\mathbf{x}}_k, \underline{\mathbf{v}}_k), \tag{5.61}$$

where $\underline{\mathbf{v}}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k)$ is the Gaussian measurement noise, where the covariance matrix $\mathbf{R}_k$ is known.

The goal is to estimate states' $\underline{\mathbf{x}}_{1:K}$ parameters using the realizations of the prior $\mathbf{x}_0$, the realizations of the interoceptive measurements $\mathbf{u}_k$ and realizations of the exteroceptive measurements $\mathbf{y}_k$. If the RVs are Gaussian, then they can be parametrized by a mean and a covariance (i.e., $\underline{\mathbf{x}}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$).

Remark 5.7.1. The interoceptive measurement $\underline{\mathbf{u}}_k$ is a random variable that can be assumed to be dis-

tributed according to

$$\underline{\mathbf{u}}_k \sim \mathcal{N}\left(\mathbf{u}_k, \mathbf{Q}_k^{\mathbf{u}}\right). \tag{5.62}$$

The measurement is assumed to be random due to measurement noise. However, the measurement noise can be embedded into the process noise $\mathbf{Q}$. Therefore, it is possible to assume that the interoceptive measurements are not random, but the interoceptive measurement process noise is indeed random and will be embedded with the process noise $\underline{\mathbf{w}}_k$.

Remark 5.7.2. The states will be assumed to be normally distributed. That is, $\underline{\mathbf{x}}_k \sim \mathcal{N}\left(\mathbf{x}_k, \mathbf{P}_k\right)$. Therefore, the parameters to be estimated should

$$\tilde{\boldsymbol{\theta}} = \left\{\mathbf{x}_0, \ldots, \mathbf{x}_N, \mathbf{P}_0, \ldots, \mathbf{P}_N\right\}. \tag{5.63}$$

However, to make things easier, the covariances $\mathbf{P}_{0:N}$ will not be estimated (they will be computed after the MAP estimate is computed). Thus, the parameters that'll be estimated is given by

$$\boldsymbol{\theta} = \left\{\mathbf{x}_0, \ldots, \mathbf{x}_N\right\}. \tag{5.64}$$

The realizations are then

$$\mathbf{X} = \left\{\check{\mathbf{x}}_0, \mathbf{u}_{0:N-1}, \mathbf{y}_{1:N}\right\}. \tag{5.65}$$

For the MAP estimator, the PDF of interest is

$$\begin{aligned}
f\left(\boldsymbol{\theta} \middle|\, \mathbf{X}\right) &= f\left(\mathbf{x}_0, \ldots, \mathbf{x}_N \middle|\, \check{\mathbf{x}}_0, \mathbf{u}_{0:N-1}, \mathbf{y}_{1:N}\right) && (5.66)\\
&= \frac{f\left(\mathbf{y}_{1:N} \middle|\, \mathbf{x}_{0:N}, \check{\mathbf{x}}_0, \mathbf{u}_{0:N-1}\right) f\left(\mathbf{x}_{0:N} \middle|\, \check{\mathbf{x}}_0, \mathbf{u}_{0:N-1}\right)}{f\left(\mathbf{y}_{1:N} \middle|\, \check{\mathbf{x}}_0, \mathbf{u}_{0:N-1}\right)} && (5.67)\\
&= \eta f\left(\mathbf{y}_1 \middle|\, \mathbf{x}_1\right) \cdots f\left(\mathbf{y}_N \middle|\, \mathbf{x}_N\right) f\left(\mathbf{x}_N \middle|\, \mathbf{x}_{N-1}, \mathbf{u}_{N-1}\right) \cdots f\left(\mathbf{x}_1 \middle|\, \mathbf{x}_0, \mathbf{u}_0\right) f\left(\mathbf{x}_0 \middle|\, \check{\mathbf{x}}_0\right) && (5.68)\\
&= \eta \prod_{k=1}^{N} f\left(\mathbf{y}_k \middle|\, \mathbf{x}_k\right) \prod_{k=1}^{N} f\left(\mathbf{x}_k \middle|\, \mathbf{x}_{k-1}, \mathbf{u}_{k-1}\right) f\left(\mathbf{x}_0, \middle|\, \check{\mathbf{x}}_0\right), && (5.69)
\end{aligned}$$

where η is a normalizing parameter that is independent of the design variables ($\mathbf{x}_{0:N}$) and the following assumptions were used

1. The measurements $\mathbf{y}_k$ depends only on $\mathbf{x}_k$ (and measurement noise $\underline{\mathbf{n}}_k$)as shown in the measurement model (5.61). Thus, the random variable $\underline{\mathbf{y}}_k$ is independent of all other random variables.
2. The state $\underline{\mathbf{x}}_k$ depends only on $\underline{\mathbf{x}}_{k-1}$ and $\mathbf{u}_{k-1}$ (and process noise $\underline{\mathbf{w}}_{k-1}$) as shown in (5.60). Therefore, the random variables $\mathbf{x}_{0:N}$ follow the **Markov sequence**: the future states $\underline{\mathbf{x}}_{k+1:N}$ are independent of the past states $\underline{\mathbf{x}}_{0:k-1}$ given the present state $\underline{\mathbf{x}}_k$. The Markov sequence is dictated by the process model (5.60).

If $\underline{\mathbf{x}}_{0:N}$ are assumed to be [marginally] Gaussian, then it is easier to minimize the negative-log of (5.66) than maximize (5.66) directly.

Taking the negative-log of (5.69) gives

$$-\log f\left(\mathbf{x}_{0:N}|\,\check{\mathbf{x}}_0, \mathbf{u}_{0:N-1}, \mathbf{y}_{1:N}\right) = \frac{1}{2}\left\|\mathbf{x}_0 - \check{\mathbf{x}}_0\right\|^2_{\check{\mathbf{P}}_0} + \frac{1}{2}\sum_{k=1}^{N}\left\|\mathbf{y}_k - \mathbf{g}(\mathbf{x}_k, \mathbf{0})\right\|^2_{\mathbf{R}_k} + \frac{1}{2}\sum_{k=1}^{N}\left\|\mathbf{x}_k - \mathbf{f}(\mathbf{x}_{k-1}, \mathbf{u}_{k-1}, \mathbf{0})\right\| + \boldsymbol{\gamma}, \tag{5.70}$$

where

$$\|\mathbf{z}\|^2_{\boldsymbol{\Sigma}} \triangleq \mathbf{z}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{z}, \tag{5.71}$$

and $\boldsymbol{\gamma}$ are constant terms that are independent of the design variables thus they will not affect the optimization.

Therefore, the MAP estimate of $\hat{\mathbf{x}}^{\text{MAP}}_{0:N}$ is the solution to

$$\hat{\mathbf{x}}^{\text{MAP}}_{0:N} = \underset{\mathbf{x}_{0:N}\in\mathbb{R}^n}{\arg\min}\, \frac{1}{2}\left\|\mathbf{x}_0 - \check{\mathbf{x}}_0\right\|^2_{\check{\mathbf{P}}_0} + \frac{1}{2}\sum_{k=1}^{N}\left\|\mathbf{y}_k - \mathbf{g}(\mathbf{x}_k, \mathbf{0})\right\|^2_{\mathbf{R}_k} + \frac{1}{2}\sum_{k=1}^{N}\left\|\mathbf{x}_k - \mathbf{f}(\mathbf{x}_{k-1}, \mathbf{u}_{k-1}, \mathbf{0})\right\|, \quad = \underset{\mathbf{x}_{0:N}\in\mathbb{R}^n}{\arg\min}\, J \tag{5.72}$$

which is a (nonlinear) weighted least squares problem.

Remark 5.7.3. In (5.70), it was assumed that $\mathbf{w}_k = \mathbf{n}_k = \mathbf{0}$. I'm not not sure why this assumption is done. The way I think of it is that the noise $\mathbf{w}_k$ is embedded into the interoceptive measurement $\mathbf{u}_k$. Similarly, the noise in $\mathbf{n}_k$ is embedded in the exteroceptive measurement $\mathbf{y}_k$.

5.7.2 Expressing the nonlinear least squares problem in matrix form

The objective function can be expressed as

$$J(\mathbf{x}_{0:N}) = \frac{1}{2}\left\|\mathbf{x}_0 - \check{\mathbf{x}}_0\right\|^2_{\check{\mathbf{P}}_0} + \frac{1}{2}\sum_{k=1}^{N}\left\|\mathbf{x}_k - \mathbf{f}(\mathbf{x}_{k-1}, \mathbf{u}_{k-1}, \mathbf{0})\right\| + \frac{1}{2}\sum_{k=1}^{N}\left\|\mathbf{y}_k - \mathbf{g}(\mathbf{x}_k, \mathbf{0})\right\|^2_{\mathbf{R}_k} \tag{5.73}$$

$$= \frac{1}{2}\mathbf{e}(\mathbf{x}_{0:N})^{\mathsf{T}}\mathbf{W}\mathbf{e}(\mathbf{x}_{0:N}), \tag{5.74}$$

where

$$\mathbf{e}(\mathbf{x}_{0:N}) = \begin{bmatrix} \mathbf{x}_0 - \check{\mathbf{x}}_0 \\ \hline \mathbf{x}_1 - \mathbf{f}\left(\mathbf{x}_0, \mathbf{u}_0, \mathbf{0}\right) \\ \vdots \\ \mathbf{x}_N - \mathbf{f}\left(\mathbf{x}_{N-1}, \mathbf{u}_{N-1}, \mathbf{0}\right) \\ \hline \mathbf{y}_1 - \mathbf{g}(\mathbf{x}_1, \mathbf{0}) \\ \vdots \\ \mathbf{y}_N - \mathbf{g}(\mathbf{x}_N, \mathbf{0}) \end{bmatrix} \tag{5.75}$$

is the *error function*, and

$$\mathbf{W}^{-1} = \begin{bmatrix} \check{\mathbf{P}}_0 & & & & & & \\ & \mathbf{Q}_0 & & & & & \\ & & \ddots & & & & \\ & & & \mathbf{Q}_N & & & \\ & & & & \mathbf{R}_1 & & \\ & & & & & \ddots & \\ & & & & & & \mathbf{R}_N \end{bmatrix} \tag{5.76}$$

is the *weight matrix*.

5.7.3 Covariance on MAP estimate

In the optimization problem above, the covariance was not estimated. However, it can be estimated by using the MAP state estimates $\hat{\mathbf{x}}_{0:N}^{\text{MAP}}$. The *joint* covariance can be approximated by

$$\text{Cov}\left[\hat{\mathbf{x}}_{0:N}^{\text{MAP}}\right] = \left(\mathbf{J}\left(\hat{\mathbf{x}}_{0:N}^{\text{MAP}}\right)^{\mathsf{T}} \mathbf{W} \mathbf{J}\left(\hat{\mathbf{x}}_{0:N}^{\text{MAP}}\right)\right)^{-1}, \tag{5.77}$$

where

$$\mathbf{J}\left(\hat{\mathbf{x}}_{0:N}^{\text{MAP}}\right) = \left.\frac{\mathrm{d}\mathbf{e}\left(\mathbf{x}_{0:N}\right)}{\mathrm{d}\mathbf{x}_{0:N}}\right|_{\mathbf{x}_{0:N}=\hat{\mathbf{x}}_{0:N}^{\text{MAP}}}. \tag{5.78}$$

is the Jacobian of the error function (5.75) w.r.t. the design variables, evaluated at the MAP estimate $\hat{\mathbf{x}}_{0:N}^{\text{MAP}}$.

Appendices

Appendix A

Linear algebra

A.1 Schur complement

Theorem A.1.1 (Schur's complement)**.** Let

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{12} & \mathbf{A}_{22} \end{bmatrix} \tag{A.1}$$

be a square matrix. If $\mathbf{A}_{11}$ is nonsingular, then the Schur complement of the block $\mathbf{A}_{11}$ of the matrix $\mathbf{A}$ is the matrix

$$\mathbf{A}/\mathbf{A}_{11} := \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}. \tag{A.2}$$

Similarly, if $\mathbf{A}_{22}$ is nonsingular, then the Schur complement of the block $\mathbf{A}_{22}$ of the matrix $\mathbf{A}$ is the matrix

$$\mathbf{A}/\mathbf{A}_{22} := \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}. \tag{A.3}$$

A.1.1 Application to solving linear equations

Given a system of equations

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}. \tag{A.4}$$

Then what is the solution $\mathbf{x}_1$ without computing $\mathbf{x}_2$ (or without solving the big system of equations all at once)?

The solution is given by using the Schur complement of the block $\mathbf{A}_{22}$ of the matrix $\mathbf{A}$. Specifically, pre-multiply the left and right sides of (A.4) by the matrix

$$\mathbf{L} = \begin{bmatrix} \mathbf{1} & -\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{0} & \mathbf{1} \end{bmatrix}. \tag{A.5}$$

Then, the result is

$$\begin{bmatrix} \mathbf{1} & -\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{1} & -\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} \tag{A.6}$$

$$\begin{bmatrix} \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{0} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{b}_2 \\ \mathbf{b}_2 \end{bmatrix}. \tag{A.7}$$

The solution $\mathbf{x}_1$ is then given by solving

$$\left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)\mathbf{x}_1 = \mathbf{b}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{b}_2. \tag{A.8}$$

A.1.2 Application in probability

A.1.2.1 Conditioning using covariance matrix

Given a the random variable

$$\begin{bmatrix} \underline{\mathbf{x}}_1 \\ \underline{\mathbf{x}}_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \bar{\mathbf{x}}_1 \\ \bar{\mathbf{x}}_2 \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{12}^{\mathsf{T}} & \mathbf{\Sigma}_{22} \end{bmatrix} \right). \tag{A.9}$$

Then, the marginal covariance of $\underline{\mathbf{x}}_1$ is $\mathbf{\Sigma}_{11}$. However, the covariance of $\underline{\mathbf{x}}_1$ *given* $\underline{\mathbf{x}}_2 = \hat{\mathbf{x}}_2$ is given by the Schur complement. Specifically,

$$\mathbb{E}\left[\underline{\mathbf{x}}_1 | \underline{\mathbf{x}}_2 = \hat{\mathbf{x}}_2\right] = \bar{\mathbf{x}}_1 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\left(\hat{\mathbf{x}}_2 - \bar{\mathbf{x}}_2\right), \tag{A.10}$$

$$\mathrm{Cov}\left[\underline{\mathbf{x}}_1 | \underline{\mathbf{x}}_2 = \hat{\mathbf{x}}_2\right] = \mathbf{\Sigma}/\mathbf{\Sigma}_{22} \tag{A.11}$$

$$= \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{12}^{\mathsf{T}}. \tag{A.12}$$

Note that this is equivalent to solving the linear system of equations

$$\begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{12}^{\mathsf{T}} & \mathbf{\Sigma}_{22} \end{bmatrix} \begin{bmatrix} \mathbb{E}\left[\underline{\mathbf{x}}_1 | \underline{\mathbf{x}}_2 = \hat{\mathbf{x}}_2\right] \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{x}}_1 \\ \bar{\mathbf{x}}_2 - \hat{\mathbf{x}}_2 \end{bmatrix}. \tag{A.13}$$

I'm not sure what's the relation between (A.13) and (A.10).

A.1.2.2 Conditioning using information matrix

On the other hand, if we are given the same random variable

$$\begin{bmatrix} \underline{\mathbf{x}}_1 \\ \underline{\mathbf{x}}_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \bar{\mathbf{x}}_1 \\ \bar{\mathbf{x}}_2 \end{bmatrix}, \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{12}^{\mathsf{T}} & \mathbf{A}_{22} \end{bmatrix}^{-1} \right), \tag{A.14}$$

where

$$\mathbf{A} = \mathbf{\Sigma}^{-1} \tag{A.15}$$

is the *information matrix*. Then, conditioning is easily done by partitioning the system. That is,

$$\mathrm{Cov}\left[\underline{\mathbf{x}}_1|\underline{\mathbf{x}}_2 = \hat{\mathbf{x}}_2\right] = \mathbf{A}_{11}^{-1}. \tag{A.16}$$

On the other hand, marginalization is more "difficult" and is given by the Schur complement. Specifically,

$$\mathrm{Cov}\left[\underline{\mathbf{x}}_1\right] = \left(\mathbf{A}/\mathbf{A}_{22}\right)^{-1} \tag{A.17}$$

$$= \left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{12}^{\mathsf{T}}\right)^{-1}. \tag{A.18}$$

Appendix B

Numerically sampling from a normal distribution

Say we want to sample realizations of a normal random variable $\underline{\mathbf{x}} \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma})$, given the mean $\boldsymbol{\mu}_x$ and covariance $\boldsymbol{\Sigma}_x$. How can this be done using MATLAB's `randn` function?

B.1 MATLAB's `randn` function

The `randn` function produces a realization of a normal random variable with zero mean and unit variance. How to compute the variance of such variable? Well, that depends on how the realizations are stored in MATLAB. That is, the realizations can be stored as columns or as rows.

Say the random variable has n degrees of freedom and we want m realizations. Thus, if the realizations are stored as columns, then the matrix is of size $n \times m$. On the other hand, if the realizations are stored as rows, then the matrix of size $m \times n$. Specifically,

$$\mathbf{x}_{\text{unit,col}} = \texttt{randn(n,m)} \in \mathbb{R}^{n \times m}, \tag{B.1}$$

$$\mathbf{x}_{\text{unit,row}} = \texttt{randn(m,n)} \in \mathbb{R}^{m \times n}. \tag{B.2}$$

The covariance of $\mathbf{x}_{\text{unit}}$ can be approximated numerically as

$$\begin{aligned} \text{Cov}\left[\mathbf{x}_{\text{unit}}\right] &= \boldsymbol{\Sigma}_{\text{unit}} && \text{(B.3)} \\ &= \mathbf{1}_{n \times n} && \text{(B.4)} \\ &\approx \frac{1}{m-1} \mathbf{x}_{\text{unit,col}} \mathbf{x}_{\text{unit,col}}^{\mathsf{T}} && \text{(B.5)} \\ &= \frac{1}{m-1} \mathbf{x}_{\text{row,unit}}^{\mathsf{T}} \mathbf{x}_{\text{row,unit}}. && \text{(B.6)} \end{aligned}$$

Well, what if we want a realization of a normal random variable with zero mean but non-unit variance $\boldsymbol{\Sigma}_x$?

The covariance is then

$$\begin{aligned}
\mathrm{Cov}\left[\mathbf{x}\right] &= \boldsymbol{\Sigma}_x && \text{(B.7)}\\
&= \mathbf{L}\underbrace{\boldsymbol{\Sigma}_{\mathrm{unit}}}_{\mathbf{I}}\mathbf{L}^{\mathsf{T}} && \text{(B.8)}\\
&= \mathbf{R}^{\mathsf{T}}\underbrace{\boldsymbol{\Sigma}_{\mathrm{unit}}}_{\mathbf{I}}\mathbf{R}, && \text{(B.9)}
\end{aligned}$$

where $\mathbf{L}$ and $\mathbf{R}$ are the lower- and upper-triangular matrices obtained from the Cholesky factorization. So which of these two should we use? Well, the answer depends on how the realizations of $\underline{\mathbf{x}}$ are stored.

If the realizations are stored in columns, then from (B.5),

$$\begin{aligned}
\mathrm{Cov}\left[\mathbf{x}\right] &= \boldsymbol{\Sigma}_x && \text{(B.10)}\\
&= \mathbf{L}\boldsymbol{\Sigma}_{\mathrm{unit}}\mathbf{L}^{\mathsf{T}} && \text{(B.11)}\\
&\approx \mathbf{L}\mathbf{x}_{\mathrm{unit,col}}\mathbf{x}_{\mathrm{unit,col}}^{\mathsf{T}}\mathbf{L}^{\mathsf{T}} && \text{(B.12)}\\
&= \underbrace{\left(\mathbf{L}\mathbf{x}_{\mathrm{unit,col}}\right)}_{\mathbf{x}_{\mathrm{col}}}\left(\mathbf{L}\mathbf{x}_{\mathrm{unit,col}}\right)^{\mathsf{T}} && \text{(B.13)}\\
&= \mathbf{x}_{\mathrm{col}}\mathbf{x}_{\mathrm{col}}^{\mathsf{T}}, && \text{(B.14)}
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{x}_{\mathrm{col}} &= \mathbf{L}\mathbf{x}_{\mathrm{unit,col}} && \text{(B.15)}\\
&= \mathbf{L}\texttt{randn(n,m)}. && \text{(B.16)}
\end{aligned}$$

It should be noted that if $\mathbf{R}$ is used in place of $\mathbf{L}$, then the results would not be the same in general. The upper triangular matrix $\mathbf{R}$ can still be used by setting

$$\mathbf{L} = \mathbf{R}^{\mathsf{T}}. \qquad \text{(B.17)}$$

Specifically,

$$\begin{aligned}
\mathrm{Cov}\left[\mathbf{x}\right] &= \boldsymbol{\Sigma}_x && \text{(B.18)}\\
&= \mathbf{R}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathrm{unit}}\mathbf{R} && \text{(B.19)}\\
&\approx \mathbf{R}^{\mathsf{T}}\mathbf{x}_{\mathrm{unit,row}}^{\mathsf{T}}\mathbf{x}_{\mathrm{unit,row}}\mathbf{R} && \text{(B.20)}\\
&= \left(\mathbf{x}_{\mathrm{unit,row}}\mathbf{R}\right)^{\mathsf{T}}\underbrace{\left(\mathbf{x}_{\mathrm{unit,row}}\mathbf{R}\right)}_{\mathbf{x}_{\mathrm{row}}} && \text{(B.21)}\\
&= \mathbf{x}_{\mathrm{row}}^{\mathsf{T}}\mathbf{x}_{\mathrm{row}}, && \text{(B.22)}
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{x}_{\mathrm{row}} &= \mathbf{x}_{\mathrm{unit,row}}\mathbf{R} && \text{(B.23)}\\
&= \texttt{randn(m,n)}\,\mathbf{R}. && \text{(B.24)}
\end{aligned}$$

Summary:

$$\begin{aligned}
\boldsymbol{\Sigma}_x &= \mathbf{R}^\mathsf{T}\mathbf{R} && \text{(B.25)}\\
&= \mathbf{L}\mathbf{L}^\mathsf{T}, && \text{(B.26)}\\
\mathbf{x}_{\text{unit,col}} &= \texttt{randn(n,m)} && \text{(B.27)}\\
&= \mathbf{x}_{\text{unit,row}}^\mathsf{T}, && \text{(B.28)}\\
\mathbf{x}_{\text{col}} &= \mathbf{L}\mathbf{x}_{\text{unit,col}}, && \text{(B.29)}\\
&= \mathbf{x}_{\text{row}}^\mathsf{T}, && \text{(B.30)}\\
\mathbf{x}_{\text{row}} &= \mathbf{x}_{\text{unit,row}}\mathbf{R}. && \text{(B.31)}
\end{aligned}$$

B.2 Sampling (another derivation)

Assume we have N realizations $\mathbf{x} \in \mathbb{R}^{n\times N}$ of the random variable $\underline{\mathbf{x}} \in \mathbb{R}^n$. For simplicity, we'll assume the random variable has zero mean. Let's assume that there exists a matrix $\mathbf{R}$ such that

$$\underline{\mathbf{x}} = \mathbf{R}\underline{\mathbf{x}}_0, \tag{B.32}$$

where $\underline{\mathbf{x}}_0 \sim N(\mathbf{0}, \mathbf{1})$. Then,

$$\begin{aligned}
\boldsymbol{\Sigma} &= \mathbb{E}\left[\underline{\mathbf{x}}\underline{\mathbf{x}}^\mathsf{T}\right] && \text{(B.33)}\\
&= \mathbf{R}\mathbb{E}\left[\underline{\mathbf{x}}_0\underline{\mathbf{x}}_0^\mathsf{T}\right]\mathbf{R}^\mathsf{T} && \text{(B.34)}\\
&= \mathbf{R}\mathbf{R}^\mathsf{T}. && \text{(B.35)}
\end{aligned}$$

well, how can we construct $\mathbf{R}$? Decompose the covariance matrix using eigendecomposition. Specifically,

$$\begin{aligned}
\boldsymbol{\Sigma} &= \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\mathsf{T} && \text{(B.36)}\\
&= \underbrace{\mathbf{U}\sqrt{\boldsymbol{\Lambda}}}_{\mathbf{R}}\mathbf{1}\underbrace{\sqrt{\boldsymbol{\Lambda}}\mathbf{U}^\mathsf{T}}_{\mathbf{R}^\mathsf{T}} && \text{(B.37)}\\
&= \mathbf{R}\mathbf{R}^\mathsf{T}, && \text{(B.38)}
\end{aligned}$$

where $\mathbf{U}$ and $\boldsymbol{\Lambda}$ are the eigenvector and the eigenvalue matrices, respectively. Thus,

$$\mathbf{R} = \mathbf{U}\sqrt{\boldsymbol{\Lambda}}. \tag{B.39}$$

Therefore, the random variable $\underline{\mathbf{x}}$ can be sampled by sampling from the unit normal random variable $\underline{\mathbf{x}}_0$ and scaling it using

$$\mathbf{x} = \mathbf{U}\sqrt{\boldsymbol{\Lambda}}\mathbf{x}_0. \tag{B.40}$$

B.3 Covariance ellipses

To plot covariance ellipses (in 2D), eigendecomposition is used. Specifically, given a covariance matrix $\mathbf{\Sigma}$, then decompose

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathsf{T}}, \tag{B.41}$$

where $\mathbf{U}^{-1} = \mathbf{U}^{\mathsf{T}}$ because $\mathbf{\Sigma} = \mathbf{\Sigma}^{\mathsf{T}}$ is symmetric. The standard deviation ellipses[1] are the computed as

$$\boldsymbol{\sigma} = \mathbf{U}\sqrt{\mathbf{\Lambda}} \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix}, \qquad \theta \in [0, 2\pi). \tag{B.42}$$

Figure B.1 shows an example of such ellipses.

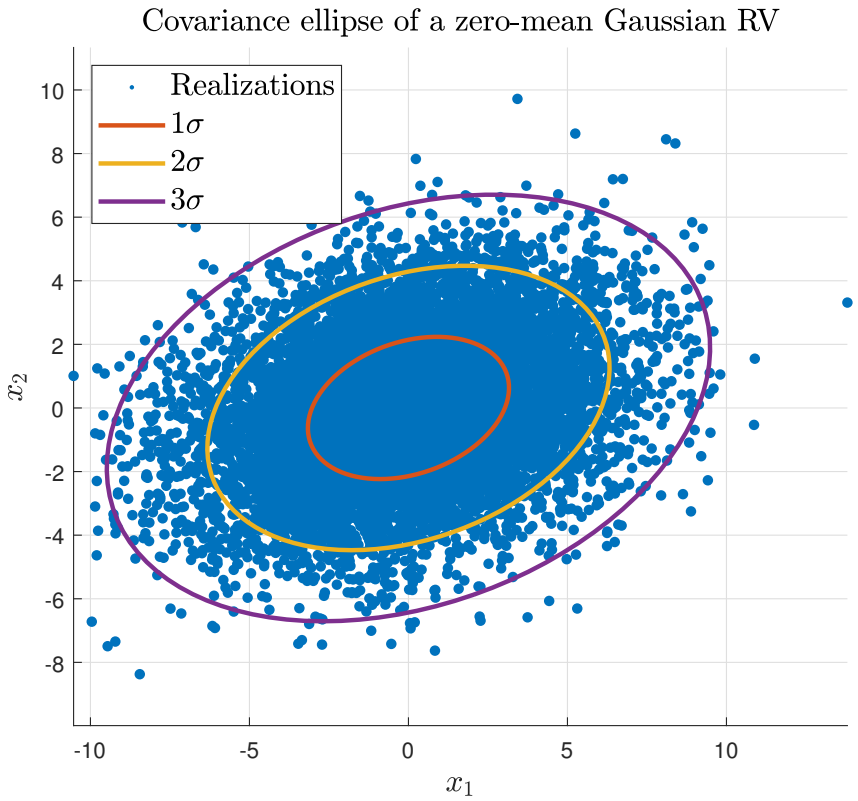

Figure B.1: Covariance ellipses (actually, standard deviation ellipses) around $N = 10^4$ samples.

[1] I'm not sure about the naming here. But the ellipse computed represents the 1σ bounds.

Bibliography

[1] I. Psaromiligkos, "Slides from ECSE 509: Probability and random signals 2 taught in fall 2019 at McGill university."

[2] K. P. Murphy, *Machine learning: a probabilistic perspective*.

[3] F. Gustafsson and G. Hendeby, "On nonlinear transformations of stochastic variables and its application to nonlinear filtering," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 3617–3620, ISSN: 1520-6149. [Online]. Available: http://ieeexplore.ieee.org/document/4518435/

[4] S. Sarkka, *Bayesian Filtering and Smoothing*. Cambridge University Press. [Online]. Available: http://ebooks.cambridge.org/ref/id/CBO9781139344203

[5] T. D. Barfoot, *State Estimation for Robotics: A Matrix Lie Group Approach*. Cambridge University Press.